Centre de recherche sur les transports (C.R.T.)
Centre for Research on Transportation

CRT-2003-20

ANALYSE DU DEGRÉ D'ASSOCIATION ENTRE L'USAGE DU TÉLÉPHONE MOBILE PENDANT LA CONDUITE ET LES ACCIDENTS DE VOITURE

par

Stéphane Courchesne

Mémoire de maîtrise

Septembre 2003

Université de Montréal
C.P. 6128, succursale Centre-ville, Montréal QC H3C 3J7 Canada
Téléphone : (514) 343-7575 / Télécopieur : (514) 343-7121

crt@crt.umontreal.ca
www.crt.umontreal.ca

HEC ÉCOLE POLYTECHNIQUE MONTRÉAL Université de Montréal

104-00215/

PUBLICATION
CRT-2003-20

ANALYSE DU DEGRÉ D'ASSOCIATION ENTRE L'USAGE DU TÉLÉPHONE MOBILE PENDANT LA CONDUITE ET LES ACCIDENTS DE VOITURE

par

Stéphane Courchesne [1]

Mémoire de maîtrise présenté en vue de l'obtention du diplôme de M.Sc. en statistique.
Directeur de recherche : Jean-François Angers; codirecteur : François Bellavance.

[1] Département de mathématiques et de statistique, et Centre de recherche sur les transports, Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal, Canada H3C 3J7

Dépôt légal – Bibliothèque nationale du Québec
Bibliothèque nationale du Canada, 2003

SOMMAIRE

L'avancement de la technologie dans le secteur des communications sans fil a grandement favorisé l'usage du téléphone mobile/cellulaire. Depuis quelques années, le phénomène de l'utilisation du téléphone mobile pendant la conduite suscite beaucoup de questions. Doit-on légiférer l'usage de cet appareil dans une voiture ? Contribue-t-il à augmenter le risque d'avoir un accident de la route ?

Dans ce mémoire, nous tentons de fournir un argument statistique qui permettrait de conclure sur l'impact qu'a l'utilisation du téléphone mobile pendant la conduite. Pour ce faire, nous allons estimer deux mesures d'association entre les facteurs "faire l'usage d'un téléphone mobile en conduisant" et "avoir un accident de voiture". Ces mesures sont le rapport de cotes et le risque relatif et sont obtenues à partir d'un tableau de contingence 2×2.

Essentiellement, nous allons développer une méthode pour reconstruire un tableau de contingence. Nous estimerons dans un premier temps la densité de la répartition des appels faits ou reçus, de même que la densité de la répartition des accidents de la route. Nous emploierons l'algorithme EM afin d'obtenir l'estimation des paramètres qui ajustent le mieux une densité aux données. Grâce aux densités obtenues pour les appels et pour les accidents, nous pouvons obtenir un histogramme des probabilités de faire usage de son téléphone mobile au moment où un accident de la route survient ou encore d'être impliqué dans un accident à ce même moment.

À l'aide de l'ajustement d'une densité bêta à chacun des histogrammes obtenus, nous pouvons reconstruire le tableau de contingence 2×2 et estimer le rapport de cotes ainsi que le risque relatif grâce à la méthode de Monte Carlo avec fonction d'importance. Nous obtenons également des régions α-crédibles pour chacune des mesures d'association.

Cette méthode a été appliquée à quatre groupes d'intérêt soient :

- tous les seuls usagers de téléphone cellulaire avec tous les types d'accidents,
- tous les seuls usagers de téléphone cellulaire avec les accidents impliquant des blessés,
- tous les seuls grands utilisateurs de téléphone cellulaire (moyenne d'au moins sept appels par jour) avec tous les types d'accidents,
- tous les seuls petits utilisateurs de téléphone cellulaire (moyenne inférieure à un appel par jour) avec tous les types d'accidents.

De façon générale, nous remarquons que le risque relatif augmente significativement pour tous les groupes à l'exception des petits utilisateurs. Plus précisément, pour le groupe des seuls utilisateurs avec tous les types d'accidents, le risque relatif se situe à 1,74. Ce risque passe à 2,27 pour le groupe des seuls grands utilisateurs (moyenne d'au moins sept appels par jour) pour tous les types d'accidents. Si nous nous limitons aux accidents impliquant des blessés pour le groupe des seuls utilisateurs, le risque devient 2,00. Enfin, les petits utilisateurs d'un téléphone mobile (pour tous les types d'accidents) ont un risque relatif qui se situe à 0,837.

Mots-clés : Tableau de contingence, information manquante, algorithme EM, Monte-Carlo, statistique bayésienne.

SUMMARY

The advance of technology in the sector of the communications has largely favored the use of the mobile/cellular phone. For a few years, many questions arise regarding the phenomenon of the use of the mobile phone while driving. Does one have to legislate the use of the phone in a car? Does it contribute to increase the risk to have a car accident?

In this work, we try to provide statistical arguments which would make it possible to conclude on the impact that the use of the mobile phone has while driving. To achieve this goal, we will estimate two measurements of association between the factors "to use a mobile phone while driving" and "to have a car accident". These measures are the odds ratio and the relative risk. They are obtained from a 2×2 contingency table.

We will essentially develop a method to rebuild a contingency table. We will first estimate the density of the distribution of making or receiving a call, as well as the density of the distribution of having a car accident. We will use the EM algorithm in order to obtain the estimate of the parameters of a density which best adjust the data. With the densities obtained for both the calls and the accidents, we can obtain a histogram of the probabilities of using a mobile telephone at the time when a car accident occurs or being implied in an accident at this same time.

Using the beta density obtained from each histogram, we can rebuild the 2×2 contingency table and estimate the odds ratio as well as the relative risk using the Monte Carlo importance sampling method. We also obtain α-credible set for each measure of association.

This method was applied to four groups of interest :

- all unique mobile phone users for all types of accidents,
- all unique mobile phone users for accidents involving injuries,
- all unique mobile phone users with an average of at least seven calls per day for all types of accidents,
- all unique mobile phone users with an average of less than a call per day for all types of accidents.

In a general way, we notice that the relative risk increases significantly for all the groups except for the small users. More precisely, for the group of the only users with all the types of accidents, the relative risk is 1,74. This risk passes to 2,27 for the group of the large only users (average of at least seven calls per day) for all the types of accidents. If we limit ourselves to the accidents implying of wounded for the group of the only users, the risk becomes 2,00. Lastly, the small users of a mobile telephone (for all the types of accidents) have a relative risk which is 0,837.

Key words : Contingency table, missing information, EM algorithm, Monte Carlo, Bayesian Statistics.

REMERCIEMENTS

Je tiens d'abord à remercier les gens du Centre de recherche sur les transports, tout particulièrement François Bellavance pour m'avoir offert cette opportunité de recherche sur un sujet très actuel et le Dre Claire Laberge-Nadeau qui m'a permis de réaliser ce projet. Je voudrais ensuite souligner l'aide précieuse que m'a fournie Denise Desjardins.

Comment passer sous silence le support incroyable de mon directeur de recherche Jean-François Angers. Ce travail est le résultat d'une belle expérience qui n'aurait été aussi enrichissante sans lui. Merci Jean-François pour ton temps, ta patience et ton ingéniosité. Tu sais véritablement prendre plaisir à explorer de nouvelles idées, et tu possèdes la rare qualité de bien transmettre ta passion pour les statistiques.

Finalement, je me dois de remercier ma famille, Mélanie, Priscilla, Sébastien, mes collègues de travail et mes amis pour avoir accepté, bien malgré eux, la tâche d'entendre parler de téléphones mobiles et d'accidents de voiture pendant plusieurs mois!

Table des matières

Sommaire iii

Summary v

Remerciements vii

Table des figures xi

Liste des tableaux xiv

Introduction 1

Chapitre 1. Éléments de théorie statistique 3

1.1. Données Catégorielles 3

1.1.1. Tableaux de contingence 2×2 5

1.1.2. Indépendance des variables catégorielles 10

1.1.3. Rapport de cotes 11

1.1.4. Propriétés du rapport de cotes 13

1.1.5. Le risque relatif 14

1.2. Données circulaires 16

1.2.1. Contexte circulaire 17

1.2.1.1. Contexte circulaire paramétrique 21

1.2.1.2. Distribution de von Mises 22

1.2.2. Présentation de l'exemple 24

Chapitre 2. Estimation des paramètres pour la répartition des accidents et des appels selon le moment de la journée 26

2.1. Contexte bayésien 27
2.1.1. Éléments de théorie bayésienne 27
2.1.2. Description de la méthode ML-II 32

2.2. Mélange de lois 35
2.2.1. Identifiabilité 36
2.2.2. Algorithme EM 37
2.2.2.1. Modèle augmenté 38
2.2.2.2. Étapes de l'algorithme EM 39
2.2.2.3. Algorithme EM pour un mélange de densités marginales à g composantes 40
2.2.2.4. Application de l'algorithme aux données de l'exemple 44

2.3. Sélection du modèle d'ajustement 45
2.3.1. Application des critères de sélection de modèle 48

Chapitre 3. Étude épidémiologique et résultats 50

3.1. Présentation de l'étude 50
3.1.1. Information sur le jeu de données 52

3.2. Modélisation des probabilités marginales 58
3.2.1. Données disponibles 58
3.2.2. Modélisation de la répartition des appels et des accidents 60
3.2.3. Probabilités marginales 65
3.2.3.1. Modélisation par une densité bêta modifiée 69
3.2.3.2. Application de l'algorithme EM à un mélange de densités bêta 73

3.3. Reconstruction du tableau de contingence 78
3.3.1. Méthode de Monte Carlo avec fonction d'importance 80

3.4. Application de la méthode à des groupes spécifiques 86

Conclusion 89

Annexe A. Données utilisées pour l'exemple 92

Annexe B. Résultats pour le groupe des seuls utilisateurs en considérant uniquement les accidents avec blessés ... 93

Annexe C. Résultats pour le groupe des grands utilisateurs en considérant tous les types d'accidents 99

Annexe D. Résultats pour le groupe des petits utilisateurs en considérant tous les types d'accidents 106

Annexe E. Procédure générale 112

Bibliographie 113

Table des figures

1.2.1 Illustration géométrique de $\bar{y}_t$ et R. 20

1.2.2 Histogramme comparatif des données de l'exemple et de la densité de VM(3,648 ; 0,278) 25

2.2.1 Comparaison de la répartition des données de l'exemple avec une approximation par une densité marginale et un mélange de deux densités marginales 46

3.1.1 Courbe de la taille de l'échantillon (N) en fonction de f selon l'équation (3.1.1) 54

3.2.1 Comparaison de la répartition des appels selon l'heure de la journée et son approximation par la méthode ML-II. 62

3.2.2 Comparaison de la répartition des appels selon l'heure de la journée et son approximation selon le nombre de modes. 62

3.2.3 Comparaison de la répartition des accidents selon l'heure de la journée et son approximation par la méthode ML-II. 64

3.2.4 Comparaison de la répartition des accidents selon l'heure de la journée et son approximation selon le nombre de modes. 65

3.2.5 Répartition des probabilités d'appel sur un intervalle de 15 minutes... 68

3.2.6 Répartition des probabilités d'accident sur un intervalle de 15 minutes. 69

3.2.7 Comparaison de la répartition des probabilités transformées d'appel et son approximation par le maximum de vraisemblance 74

3.2.8 Comparaison de la répartition des probabilités transformées d'accident et son approximation par le maximum de vraisemblance..... 74

3.2.9 Comparaison de la répartition des probabilités transformées d'appel et son approximation par différents nombres de composantes dans un mélange de densités bêta 77

3.2.10 Comparaison de la répartition des probabilités transformées d'accident et son approximation par différents nombres de composantes dans un mélange de densités bêta 77

B.1 Comparaison de la répartition des accidents selon l'heure de la journée et son approximation par un mélange à trois composantes. 94

B.2 Comparaison de la répartition des probabilités transformées d'appel et son approximation par un modèle à trois composantes dans un mélange de densités bêta 96

B.3 Comparaison de la répartition des probabilités transformées d'accident et son approximation par un modèle à trois composantes dans un mélange de densités bêta 97

C.1 Comparaison de la répartition des appels selon l'heure de la journée et son approximation par différents modèles. 101

C.2 Comparaison de la répartition des accidents selon l'heure de la journée et son approximation par différents modèles. 101

C.3 Comparaison de la répartition des probabilités transformées d'appel et son approximation par différents modèles 103

C.4 Comparaison de la répartition des probabilités transformées d'accident et son approximation par différents modèles 104

D.1 Comparaison de la répartition des appels selon l'heure de la journée et son approximation par différents modèles. 108

D.2 Comparaison de la répartition des accidents selon l'heure de la journée et son approximation par différents modèles. 109

D.3 Comparaison de la répartition des probabilités transformées d'appel et son approximation par différents modèles 110

D.4 Comparaison de la répartition des probabilités transformées d'accident et son approximation par différents modèles 111

Liste des tableaux

1.1.1 Tableau de contingence 2 × 2. 5

1.1.2 Tableau des probabilités conjointes, conditionnelles et marginales dans un tableau de contingence 2 × 2. 6

1.1.3 Exemple d'un tableau de contingence. 9

1.1.4 Tableau de contingence des probabilités conjointes, conditionnelles et marginales pour l'exemple sur les téléphones mobiles et les accidents.. 10

2.2.1 Estimation des paramètres d'une seule densité marginale et d'un mélange à deux composantes pour les données de l'exemple 46

2.3.1 Comparaison des différents modèles pour les données de l'exemple.... 48

3.1.1 Répartition des 175 000 titulaires de permis selon l'âge et le sexe..... 55

3.1.2 Taux de réponse selon l'âge et le sexe 55

3.1.3 Nombre de titulaires détenant un permis de conduire de classe 5 au 31 décembre de l'année en cours 56

3.1.4 Nombre moyen d'accidents annuels pour 1000 titulaires de l'échantillon selon le sexe et l'année entre 1996 et 1999. 57

3.1.5 Nombre moyen d'accidents annuels pour 1000 titulaires de la population du Québec âgés entre 16 et 64 ans selon le sexe et l'année. 57

3.1.6 Nombre d'accidents dans lequel un participant a été impliqué de novembre 1997 à novembre 1999 selon le sexe et son usage du téléphone cellulaire 58

3.2.1 Tableau de contingence 2 × 2 modifié. 59

3.2.2 Estimation des paramètres par la méthode ML-II pour les appels..... 61

3.2.3 Estimation des paramètres d'un mélange à deux ou trois composantes pour les données sur les appels..... 63

3.2.4 Comparaison des différents modèles pour les données sur les appels... 63

3.2.5 Estimation des paramètres par la méthode ML-II pour les accidents.. 64

3.2.6 Estimation des paramètres d'un mélange de densités marginales à deux ou trois composantes pour les données sur les accidents 66

3.2.7 Estimation des paramètres d'un mélange de densités marginales à quatre ou cinq composantes pour les données sur les accidents..... 67

3.2.8 Comparaison des différents modèles pour les données sur les accidents 68

3.2.9 Estimation des paramètres par le maximum de vraisemblance pour la distribution des probabilités des accidents et des appels 73

3.2.10 Estimation des paramètres d'un mélange de densités bêta à deux ou trois composantes pour les données sur les appels 76

3.2.11 Estimation des paramètres d'un mélange de densités bêta à deux ou trois composantes pour les données sur les accidents..... 76

3.2.12 Comparaison des différents modèles pour les accidents et les appels... 78

3.3.1 Estimation du rapport de cotes et du risque relatif par la méthode de Monte Carlo avec fonction d'importance pour n_{11} 84

3.3.2 Estimation du rapport de cotes et du risque relatif par la méthode de Monte Carlo avec fonction d'importance pour $n_{11}^{\min}$ 84

3.3.3 Estimation du rapport de cotes et du risque relatif par la méthode de Monte Carlo avec fonction d'importance pour $n_{11}^{\max}$ 85

3.3.4 Comparaison des régions α-crédibles pour le risque relatif et le rapport de cotes en considérant n_{11}, $n_{11}^{\min}$ et $n_{11}^{\max}$ au niveau $1-\alpha = 0{,}95$..... 86

3.4.1 Estimation du rapport de cotes et du risque relatif par la méthode de Monte Carlo avec fonction d'importance pour les seuls utilisateurs pour les accidents avec blessés 87

3.4.2 Estimation du rapport de cotes et du risque relatif par la méthode de Monte Carlo avec fonction d'importance pour les grands utilisateurs pour tous types d'accidents 87

3.4.3 Estimation du rapport de cotes et du risque relatif par la méthode de Monte Carlo avec fonction d'importance pour les petits utilisateurs pour tous types d'accidents 88

3.4.4 Comparaison des régions α-crédibles pour le risque relatif et le rapport de cotes avec l'hypothèse n_{11}, pour chacun des groupes de niveau $1 - \alpha = 0{,}95$. 88

B.1 Comparaison des différents modèles pour les données sur les accidents 93

B.2 Estimation des paramètres d'un mélange de densités marginales à trois composantes pour les données sur les accidents 94

B.3 Comparaison des différents modèles pour les probabilités transformées pour les appels 95

B.4 Estimation des paramètres d'un mélange de densités bêta à deux composantes pour les données sur les appels 95

B.5 Comparaison des différents modèles pour les probabilités transformées pour les accidents 95

B.6 Estimation des paramètres d'un mélange de densités bêta à deux composantes pour les données sur les appels 96

B.7 Estimation du rapport de cotes et du risque relatif par la méthode de Monte Carlo avec fonction d'importance pour les seuls utilisateurs pour les accidents avec blessés 97

B.8 Comparaison des régions α-crédibles pour le risque relatif et le rapport de cotes selon l'hypothèse et de niveau $1 - \alpha = 0{,}95$. 98

C.1 Comparaison des différents modèles pour les données sur les appels... 99

C.2 Comparaison des différents modèles pour les données sur les accidents 100

C.3 Estimation des paramètres de la densité unimodale pour les données sur les appels 100

C.4 Estimation des paramètres d'un mélange de densités marginales à trois composantes pour les données sur les accidents 100

C.5 Comparaison des différents modèles pour les probabilités transformées pour les appels 102

C.6 Estimation des paramètres d'un mélange de densités bêta à trois composantes pour les données sur les appels 102

C.7 Comparaison des différents modèles pour les probabilités transformées pour les accidents 102

C.8 Estimation des paramètres d'un mélange de densités bêta à deux composantes pour les données sur les accidents 103

C.9 Estimation du rapport de cotes et du risque relatif par la méthode de Monte Carlo avec fonction d'importance pour les grands utilisateurs pour tous les types d'accidents 104

C.10 Comparaison des régions α-crédibles pour le risque relatif et le rapport de cotes selon l'hypothèse et de niveau $1 - \alpha = 0{,}95$. 105

D.1 Comparaison des différents modèles pour les données sur les appels... 106

D.2 Comparaison des différents modèles pour les données sur les accidents 107

D.3 Estimation des paramètres de la densité unimodale pour les données sur les appels 107

D.4 Estimation des paramètres d'un mélange de densités marginales à quatre composantes pour les données sur les accidents 108

D.5 Comparaison des différents modèles pour les probabilités transformées pour les appels 109

D.6 Estimation des paramètres d'un mélange de densités bêta à trois composantes pour les données sur les appels 109

D.7 Comparaison des différents modèles pour les probabilités transformées pour les accidents 110

D.8 Estimation des paramètres d'un mélange de densités bêta à deux composantes pour les données sur les accidents 110

D.9 Estimation du rapport de cotes et du risque relatif par la méthode de Monte Carlo avec fonction d'importance pour les petits utilisateurs pour tous les types d'accidents 111

D.10 Région α-crédible pour le risque relatif et le rapport de cotes de niveau $1 - \alpha = 0{,}95$. 111

INTRODUCTION

La forte croissance de l'usage du téléphone mobile/cellulaire durant la conduite automobile soulève la question suivante : y a-t-il un risque plus élevé d'accidents de la route si ces appareils sont utilisés en conduisant ? Plusieurs chercheurs ont tenté de fournir un élément de réponse par diverses approches. Dans ce mémoire, nous analyserons des données recueillies lors d'une vaste enquête réalisée par le Centre de recherche sur les transports (C.R.T.) à laquelle la Société de l'assurance automobile du Québec (SAAQ) ainsi que quatre compagnies de téléphonie mobile (Bell, Cantel, Clearnet et Fido) ont participé. Nous tentons de quantifier le degré d'association, s'il existe, entre les facteurs "faire usage d'un téléphone mobile en conduisant" et "avoir un accident de voiture".

Au premier chapitre, nous présentons quelques éléments de théorie sur les tableaux de contingence 2×2. Nous voyons entres autres les deux mesures d'association que nous voulons estimer : le rapport de cotes et le risque relatif. Nous introduisons ensuite l'analyse des données circulaires puisque les données disponibles sont en fait des mesures de temps dans un intervalle d'une journée. La loi de von Mises, l'analogue de la loi normale dans un contexte circulaire, sera présentée ainsi que quelques unes de ses propriétés. Enfin, nous présentons un exemple qui illustre les données circulaires et auquel nous appliquerons par la suite la méthode développée.

Au second chapitre, nous débutons par établir les concepts de base en statistique bayésienne. Cette approche est essentiellement justifiée par le fait qu'il est possible de tenir compte et de l'information disponible *a priori* et de l'information

fournie par les données pour mieux représenter la réalité. Nous utilisons ensuite la méthode du maximum de vraisemblance de type II (ML-II) conjointement avec l'algorithme Espérance-Maximisation (EM) pour des mélanges de lois à g composantes afin d'obtenir des densités qui ajustent les données de l'exemple présenté au premier chapitre. Afin de statuer sur le modèle à conserver pour l'ajustement, nous utilisons d'abord le critère du maximum d'entropie qui ne pénalise pas selon le nombre de paramètres impliqués dans le modèle et la taille de l'échantillon. Dans l'éventualité où ce critère ne nous permettrait pas de conclure, nous utiliserons les critères AIC et BIC.

Finalement, le troisième chapitre débute par la présentation de l'étude épidémiologique réalisée par les chercheurs du C.R.T. et des données disponibles. Nous appliquons par la suite la méthode introduite au deuxième chapitre pour modéliser la répartition des appels (faits et reçus) et des accidents, selon l'heure de la journée. Nous allons utiliser les densités obtenues pour les appels ainsi que pour les accidents afin de trouver la densité des probabilités de faire usage d'un téléphone mobile au moment où un accident de voiture survient ou d'avoir un accident de la route au même moment. Nous possédons maintenant les informations nécessaires pour reconstruire un tableau de contingence. Nous allons utiliser la technique de Monte Carlo avec fonction d'importance afin d'obtenir des estimations pour le rapport de cotes et le risque relatif. Cette méthode nous permettra également de définir des régions α-crédibles avec probabilités égales dans les queues pour ces deux mesures d'association. Nous fournirons les résultats obtenus pour les groupes des seuls utilisateurs (tous les types d'accidents), seuls utilisateurs (accidents avec blessés uniquement), seuls utilisateurs faisant en moyenne au moins sept appels par jour (tous les types d'accidents) et seuls utilisateurs faisant en moyenne moins d'un appel par jour (tous les types d'accidents).

Chapitre 1

ÉLÉMENTS DE THÉORIE STATISTIQUE

Ce premier chapitre se veut d'abord une brève exposition du problème statistique qui a motivé le présent travail. Dans un premier temps, nous allons mettre en contexte ce mémoire en tant que part d'une étude globale, réalisée par une équipe de chercheurs [1] du Centre de recherche sur les transports (C.R.T.), sur les téléphones mobiles/cellulaires et le risque d'accident.

À la section suivante, nous présenterons la théorie sur les tableaux de contingence qui nous servira ultimement à estimer l'association entre les événements "avoir un accident de la route" et "faire l'usage d'un téléphone mobile/cellulaire pendant la conduite". Toutefois, comme les données disponibles sont des mesures de temps dans une journée, nous devrons utiliser une approche non pas linéaire mais circulaire. Pour terminer ce chapitre, nous allons donc exposer quelques notions de base sur l'analyse de données circulaires et présenter la loi de von Mises.

1.1. Données Catégorielles

En janvier 2001, des chercheurs du C.R.T. ont déposé des résultats liés à une vaste étude dans le but de vérifier certaines hypothèses reliées à l'usage du téléphone cellulaire et la conduite automobile (voir Laberge-Nadeau, Maag, Bellavance, Desjardins, Messier et Saïdi, 2001). Une de ces hypothèses étaient de

[1] Claire Laberge-Nadeau, Urs Maag, François Bellavance, Denise Desjardins , Stéphane Messier, Abdelnasser Saïdi

vérifier si la possession d'un téléphone cellulaire influence le risque d'avoir un accident.

Nous allons ici nous intéresser à cette hypothèse mais dans un cadre plus précis : est-ce que l'utilisation du téléphone cellulaire au volant influence le risque d'avoir un accident ? Pour ce faire, nous devons être en mesure de quantifier ce risque. Nous proposerons donc des méthodes d'inférence basées sur des variables aléatoires catégorielles qui nous permettront d'atteindre notre but. Nous définirons d'abord ce qu'est une telle variable ainsi que le type précis auquel nous sommes intéressés. Dans cette section, les définitions, les propositions et les théorèmes sont issus de Agresti (1990).

Définition 1.1.1. *Une variable catégorielle est une variable aléatoire dont l'échelle de mesure consiste en un nombre fini de modalités.*

Il existe plusieurs types de variables catégorielles : nominales, ordinales et à intervalle. Les variables nominales n'ont pas d'ordre naturel pour les différentes modalités, par exemple la variable sexe (homme ou femme). Les variables ordinales ont des modalités ordonnées comme, par exemple, le format d'une boisson gazeuse (petit, moyen ou grand). Finalement, les variables à intervalle sont celles qui ont une distance quantifiable entre les modalités comme la pression sanguine et le revenu d'un individu. Toutefois, pour les fins de notre analyse, nous limiterons notre étude aux variables nominales à deux modalités.

Définition 1.1.2. *Une variable catégorielle est dite qualitative nominale si en plus de ne de pas avoir d'ordre naturel dans les modalités, elle ne peut se mesurer.*

On peut citer comme exemple une variable telle l'usage ou non d'un téléphone mobile ou encore le fait d'avoir ou non un accident de voiture à un certain moment.

1.1.1. Tableaux de contingence 2 × 2

Supposons que nous avons n sujets à classer selon X et Y, deux variables catégorielles à deux modalités chacune. Pour chacun des sujets, nous pouvons trouver le couple (X_i, Y_i) $i = 1, \ldots, n$. Nous sommes intéressés à représenter ces couples de façon à pouvoir analyser la distribution conjointe de ces deux variables. Pour ce faire, nous utilisons un tableau de contingence (terme introduit par Karl Pearson, 1904).

Définition 1.1.3. *Un tableau de contingence 2 × 2 représente conjointement la répartition des observations selon chacune des variables. L'élément de la cellule (i,j), i,j = 1,2 est noté n_{ij}, et il représente le nombre d'observations qui ont simultanément les modalités i et j pour les variables X et Y respectivement.*

TABLEAU 1.1.1. Tableau de contingence 2 × 2.

X / Y	**Modalité 1**	**Modalité 2**	**Total**
Modalité 1	n_{11}	n_{12}	n_{1+}
Modalité 2	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

Le tableau 1.1.1 montre un tableau de contingence 2 × 2. Nous remarquons que le symbole "+", placé en indice, signifie qu'une sommation a été faite sur l'indice qu'il remplace. En somme pour un tableau de contingence 2 × 2, où chacune des variables X et Y sont des variables aléatoires catégorielles nominales à deux modalités, nous avons :

$$n_{i+} = \sum_{j=1}^{2} n_{ij}, \quad n_{+j} = \sum_{i=1}^{2} n_{ij} \text{ et } n = \sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij}.$$

Nous notons $\mathbb{P}(X = i, Y = j) = \nu_{ij}$ la probabilité que (X,Y) se retrouve à la fois à la ligne i et à la colonne j du tableau de contingence. Intuitivement, nous posons que $\nu_{ij} = n_{ij}/n$. La distribution des probabilités ν_{ij} est la distribution

conjointe de X et de Y.

TABLEAU 1.1.2. Tableau des probabilités conjointes, conditionnelles et marginales dans un tableau de contingence 2×2.

X / Y	Modalité 1	Modalité 2	Total
Modalité 1	ν_{11} $(\nu_{1\|1})$	ν_{12} $(\nu_{2\|1})$	ν_{1+} (1,0)
Modalité 2	ν_{21} $(\nu_{1\|2})$	ν_{22} $(\nu_{2\|2})$	ν_{2+} (1.0)
Total	ν_{+1}	ν_{+2}	1,0

Quand les deux variables sont des variables réponses, nous pouvons décrire leur association soit en utilisant leur probabilité conjointe, soit en trouvant la probabilité conditionnelle de X sachant Y ou encore la probabilité conditionnelle de Y étant donné X. La probabilité conditionnelle de Y étant donné X est reliée à la probabilité conjointe par :

$$\nu_{Y=j \mid X=i} = \frac{\nu_{ij}}{\nu_{i+}} \quad \forall i, j.$$

Le tableau 1.1.2 présente la répartition des probabilités conjointes, conditionnelles et marginales dans un tableau de contingence 2×2.

Nous pouvons considérer que le classement des observations consiste à répéter indépendamment n fois une épreuve qui admet 4 possibilités différentes de probabilités respectives $\nu_{11}, \nu_{12}, \nu_{21}, \nu_{22}$ et telles que

$$\sum_{i=1}^{2} \sum_{j=1}^{2} \nu_{ij} = 1.$$

Désignons par C_{ij} la variable aléatoire représentant le nombre d'épreuves ayant abouti au résultat de type (X_i, Y_j) dans la série complète des n réalisations. Nous aurons alors

$$P(C_{11}=n_{11},C_{12}=n_{12},C_{21}=n_{21},C_{22}=n_{22})=\frac{n!}{n_{11}!n_{12}!n_{21}!n_{22}!}\nu_{11}^{n_{11}}\nu_{12}^{n_{12}}\nu_{21}^{n_{21}}\nu_{22}^{n_{22}}$$

pour tous les choix tels que

$$\sum_{i=1}^{2}\sum_{j=1}^{2}n_{ij}=n.$$

Nous remarquons alors que la répartition des n épreuves suit une distribution multinomiale de paramètres $\underset{\sim}{n}$, $\underset{\sim}{\nu}$, où $\underset{\sim}{n}=(n_{11},n_{12},n_{21},n_{22})$ et $\underset{\sim}{\nu}=(\nu_{11},\nu_{12},\nu_{21},\nu_{22})$. De plus, en considérant cette densité multinomiale comme fonction du vecteur $\underset{\sim}{\nu}$, nous pouvons remarquer que cette nouvelle loi est une densité de Dirichlet de vecteur de paramètres $\underset{\sim}{a}=(a_{11},a_{12},a_{21},a_{22})$, notée $\mathcal{D}_4(\underset{\sim}{a})$. En effet,

$$\begin{aligned}P(\underset{\sim}{\nu}\mid\underset{\sim}{n})&\propto\prod_{(i,j)}\nu_{ij}^{a_{ij}-1}\\&\propto\frac{\Gamma(\sum_{(i,j)}a_{ij})}{\prod_{(i,j)}\Gamma(a_{ij})}\prod_{(i,j)}\nu_{ij}^{a_{ij}-1},\end{aligned}\tag{1.1.1}$$

où $a_{ij}=n_{ij}+1$.

Si nous sommes intéressés à estimer les paramètres de cette distribution multinomiale, nous pouvons utiliser la méthode du maximum de vraisemblance (MV). Pour ce faire, nous devons d'abord trouver la fonction de vraisemblance.

Définition 1.1.4. *Étant donné les observations x_i, $i=1,...,n$ d'une distribution conjointe $f(\underset{\sim}{x}\mid\underset{\sim}{\theta})$, nous appelons fonction de vraisemblance, notée $L(\underset{\sim}{\theta}\mid\underset{\sim}{x})$, la distribution conjointe de $\underset{\sim}{x}$ comme fonction des paramètres inconnus, où $\underset{\sim}{\theta}$ est le vecteur des paramètres.*

Pour un vecteur aléatoire de densité multinomiale par exemple, nous avons

$$L(\underset{\sim}{\nu}\mid\underset{\sim}{n})=\left(\frac{n!}{\prod_{(i,j)}n_{ij}!}\right)\prod_{(i,j)}\nu_{ij}^{n_{ij}}.$$

L'estimateur du maximum de vraisemblance (EMV) est donc le vecteur des paramètres $\underset{\sim}{\nu}$ qui maximise $L(\underset{\sim}{\nu}\mid\underset{\sim}{n})$. Dans le cas qui nous intéresse, afin de maximiser la fonction de vraisemblance de la distribution multinomiale comme fonction

des ν_{ij}, il suffit de maximiser

$$\prod_{(i,j)} \nu_{ij}^{n_{ij}},$$

puisque seul ce terme fait intervenir les paramètres. Il faut de plus respecter les contraintes

$$\nu_{ij} \geq 0$$

et

$$\sum_{(i,j)} \nu_{ij} = 1. \tag{1.1.2}$$

Pour simplifier les calculs, nous choisirons généralement de maximiser le logarithme de $L(\underset{\sim}{\nu} \mid \underset{\sim}{n})$. Les valeurs des paramètres qui maximiseront $L(\underset{\sim}{\nu} \mid \underset{\sim}{n})$ vont également maximiser $\log L(\underset{\sim}{\nu} \mid \underset{\sim}{n})$, car la fonction logarithme est une fonction monotone croissante. Toujours pour la distribution multinomiale, nous avons

$$\log L(\underset{\sim}{\nu} \mid \underset{\sim}{n}) \propto \sum_{(i,j)} n_{ij} \log(\nu_{ij}).$$

Remarquer que nous devons maximiser uniquement trois paramètres (ν_{11} , ν_{12} , ν_{21} par exemple) car, par la deuxième contrainte (voir équation (1.1.2)), nous pouvons déduire la valeur du quatrième paramètre (ν_{22} dans ce cas) qui est donnée par $1 - (\nu_{11} + \nu_{12} + \nu_{21})$. Pour maximiser $\log L(\underset{\sim}{\nu} \mid \underset{\sim}{n})$ par rapport à un paramètre ν_{ij}, nous devons trouver $\hat{\nu}_{ij}$ qui sera solution de

$$\begin{aligned} &\frac{\partial \log L(\underset{\sim}{\nu} \mid \underset{\sim}{n})}{\partial \nu_{ij}} = 0 \qquad , \\ \Longleftrightarrow\ &\frac{\partial}{\partial \nu_{ij}}[n_{11}\log(\nu_{11}) + n_{12}\log(\nu_{12}) + n_{21}\log(\nu_{21}) \\ &\qquad\qquad + n_{22}\log(1 - \nu_{11} - \nu_{12} - \nu_{21})] = 0, \end{aligned}$$

pour $(i,j) \in \{(1,1),(1,2),(2,1)\}$.

Pour (i,j) fixé et tel que $(i,j) \neq (2,2)$, nous obtenons alors

$$\frac{\partial \log L(\nu \mid \underline{n})}{\partial \nu_{ij}} = \frac{n_{ij}}{\nu_{ij}} - \frac{n_{22}}{\nu_{22}} = 0.$$
$$\Longleftrightarrow \nu_{ij} = \frac{\nu_{22} n_{ij}}{n_{22}}.$$

En prenant la somme sur les (i,j) de chaque côté et en respectant l'équation (1.1.2), il en résulte que

$$\sum_{(i,j)} \nu_{ij} = \frac{\nu_{22}}{n_{22}} \sum_{(i,j)} n_{ij}$$
$$1 = \frac{n\nu_{22}}{n_{22}},$$

ce qui implique donc que $\hat{\nu}_{22} = n_{22}/n$ et de façon générale, nous obtenons

$$\hat{\nu}_{ij} = \frac{n_{ij}}{n} \quad \forall i,j = 1,2.$$

En somme, les estimateurs du maximum de vraisemblance $\hat{\nu}_{ij}$ pour les ν_{ij} dans une distribution multinomiale sont simplement les proportions échantillonnales.

TABLEAU 1.1.3. Exemple d'un tableau de contingence.

X / Y	**Accident**	**Non accident**	**Total**
Téléphone	350	1000	1350
Pas de téléphone	150	500	650
Total	500	1500	2000

Si, par exemple, nous avions à considérer le tableau de contingence 1.1.3 où les variables aléatoires catégorielles sont respectivement "avoir eu un accident de voiture au cours des 12 derniers mois" et "posséder un téléphone mobile". Nous estimerions la probabilité d'avoir eu un accident de voiture au cours des 12 derniers mois et de posséder de téléphone mobile par $\hat{\nu}_{11} = n_{11}/n = 350/2000 = 0{,}175$. De même, nous trouvons $\hat{\nu}_{12} = 0{,}5$, $\hat{\nu}_{21} = 0{,}075$, $\hat{\nu}_{22} = 0{,}25$. Nous pouvons aussi calculer la probabilité qu'un individu ait un accident sachant qu'il possédait un téléphone par $\hat{\nu}_{1|1} = \hat{\nu}_{11}/\hat{\nu}_{1+} = 0{,}175/0{,}675 = 0{,}259$. Par le même raisonnement,

TABLEAU 1.1.4. Tableau de contingence des probabilités conjointes, conditionnelles et marginales pour l'exemple sur les téléphones mobiles et les accidents.

X / Y	Accident	Non accident	Total
Téléphone	0,175 (0,259)	0,5 (0,741)	0,675 (1,0)
Pas de téléphone	0,075 (0,231)	0,25 (0,769)	0,325 (1,0)
Total	0,25	0,75	1,0

nous obtenons $\hat{\nu}_{2|1} = 0{,}741$, $\hat{\nu}_{1|2} = 0{,}231$ et $\hat{\nu}_{2|2} = 0{,}769$. Le tableau 1.1.4 représente le tableau des probabilités conjointes, conditionnelles et marginales reliées à cet exemple.

1.1.2. Indépendance des variables catégorielles

Il arrive souvent en pratique que nous sommes intéressés à tester l'indépendance de deux variables aléatoires. Nous allons ici appliquer la théorie de l'indépendance de deux variables au cadre des variables catégorielles et des tableaux de contingence 2×2.

Définition 1.1.5. *Deux variables catégorielles sont dites stochastiquement indépendantes si toutes les probabilités conjointes sont égales au produit de leurs probabilités marginales, c'est-à-dire si*

$$\nu_{ij} = \nu_{i+}\nu_{+j} \qquad \forall i, j = 1, 2.$$

En considérant les données de l'exemple fournies dans le tableau 1.1.4, nous pouvons vérifier si les deux variables aléatoires catégorielles sont indépendantes. En effet, il suffit que $\nu_{ij} = \nu_{i+}\nu_{+j} \quad \forall i, j = 1, 2$. Si nous fixons $(i, j) = (1, 1)$, nous avons $\nu_{11} = 0{,}175 \neq 0{,}16875 = \nu_{1+}\nu_{+1}$. Donc ces deux variables aléatoires catégorielles ne sont pas indépendantes, car elles ne vérifient pas la définition 1.1.5.

Dans la prochaine sous-section, plutôt que de vérifier si deux variables aléatoires catégorielles sont indépendantes, nous allons quantifier leur degré d'association.

1.1.3. Rapport de cotes

Dans un tableau de contingence 2×2, pour une rangée i, nous définissons la cote que la réponse soit dans la première colonne plutôt que dans la deuxième colonne par le rapport des probabilités conjointes, c'est-à-dire

$$\Omega_i = \nu_{i1}/\nu_{i2} \quad i = 1, 2.$$

Chaque Ω_i est non négatif puisque les probabilités ν_{i1} et ν_{i2} sont non négatives. De plus, une valeur de Ω_i supérieure à 1 signifie que la modalité 1 est plus probable que la modalité 2. Nous sommes maintenant en mesure de définir le rapport de cotes.

Définition 1.1.6. *Dans un tableau de contingence 2×2, le rapport de cotes ou ratio des produits croisés, est noté Λ, et est donné par*

$$\Lambda = \frac{\Omega_1}{\Omega_2} = \frac{\nu_{11}\nu_{22}}{\nu_{12}\nu_{21}}.$$

Le rapport de cote est aussi un nombre non négatif. Si une des probabilités vaut 0, alors le rapport de cote vaut soit 0 ou ∞. De plus si $1 < \Lambda < \infty$, les sujets de la première modalité de la variable X ont plus de chance de se retrouver dans la première modalité de la variable Y que ceux de la deuxième modalité de la variable X. À l'inverse, si $0 < \Lambda < 1$, les sujets de la deuxième modalité de la variable X ont plus de chance de se retrouver dans la première modalité de la variable Y que ceux de la première modalité de la variable X. Il est intéressant de remarquer que Λ est invariant à l'orientation du tableau, c'est-à-dire si nous changeons les lignes pour les colonnes et les colonnes pour les lignes.

Par cette définition, nous pouvons déduire le théorème suivant pour les tableaux de contingence.

Théorème 1.1.1. *Dans un tableau de contingence 2 × 2, les variables aléatoires catégorielles sont dites indépendantes si et seulement si* $\Lambda = 1$.

DÉMONSTRATION. ($\Longrightarrow$)
Comme les deux variables sont indépendantes, $\nu_{ij} = \nu_{i+}\nu_{+j}$. Si nous calculons le rapport de cotes, nous obtenons :

$$\begin{aligned}\Lambda = \frac{\Omega_1}{\Omega_2} &= \frac{\nu_{11}\nu_{22}}{\nu_{12}\nu_{21}} \\ &= \frac{(\nu_{1+}\nu_{+1})(\nu_{2+}\nu_{+2})}{(\nu_{1+}\nu_{+2})(\nu_{2+}\nu_{+1})} \\ &= 1.\end{aligned}$$

($\Longleftarrow$) Comme $\Lambda = 1$, nous avons :

$$\begin{aligned}1 &= \frac{\nu_{11}\nu_{22}}{\nu_{12}\nu_{21}} \\ \Leftrightarrow \nu_{12}\nu_{21} &= \nu_{11}\nu_{22}.\end{aligned}$$

De façon générale, nous obtenons

$$\nu_{ij}\nu_{kl} = \nu_{il}\nu_{jk}.$$

Nous pouvons alors écrire :

$$\begin{aligned}\nu_{i+}\nu_{+j} = \sum_{l=1}^{2}\nu_{il}\sum_{k=1}^{2}\nu_{kj} &= \sum_{k=1}^{2}\sum_{l=1}^{2}\nu_{il}\nu_{jk} \\ = \sum_{k=1}^{2}\sum_{l=1}^{2}\nu_{ij}\nu_{kl} &= \nu_{ij}\sum_{k=1}^{2}\sum_{l=1}^{2}\nu_{kl} \\ = \nu_{ij}.\end{aligned}$$

□

Si nous considérons le tableau de contingence 1.1.4, on trouve les valeurs $\Omega_1 = 0{,}175/0{,}5 = 0{,}35$, $\Omega_2 = 0{,}075/0{,}25 = 0{,}3$ et $\Lambda = 0{,}35/0{,}3 = 1{,}167$. Ainsi pour cet exemple, nous pouvons conclure que la deuxième modalité de la variable "avoir un accident" est plus probable que la première modalité car les $\Omega_i < 1$ pour

$i = 1, 2$. Ainsi il serait plus probable de ne pas avoir d'accident de voiture peu importe si un individu possède ou non un téléphone mobile si les résultats étaient ceux du tableau 1.1.4. De plus, comme la valeur du rapport de cotes est supérieure à 1, nous pourrions affirmer que les individus possédant un téléphone mobile ont eu plus d'accidents au cours des 12 derniers mois que ceux ne possédant pas de téléphone mobile. Il y aurait donc, selon ces résultats fictifs, une association entre ces deux variables catégorielles.

Maintenant que nous sommes en mesure de calculer une mesure d'association entre deux variables aléatoires catégorielles à l'aide du rapport de cotes, nous allons nous intéresser à certaines propriétés de cette quantité.

1.1.4. Propriétés du rapport de cotes

Comme nous avons vu précédemment, les estimateurs du maximum de vraisemblance pour les paramètres ν_{ij} provenant d'une distribution multinomiale sont donnés par

$$\hat{\nu}_{ij} = \frac{n_{ij}}{n}.$$

Nous pouvons estimer le rapport de cotes à l'aide du maximum de vraisemblance de Λ,

$$\hat{\Lambda} = \frac{\hat{\nu}_{11}\hat{\nu}_{22}}{\hat{\nu}_{12}\hat{\nu}_{21}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}.$$

Comme il est possible que le rapport de cotes soit égal à 0 ou ∞, il est impossible de calculer explicitement l'espérance et la variance de Λ. Toutefois, Haldane (1955) et Gart et Zweiful (1967) ont montré qu'en considérant plutôt l'estimateur du rapport de cotes modifié suivant

$$\tilde{\Lambda} = \frac{(n_{11} + 0{,}5)(n_{22} + 0{,}5)}{(n_{12} + 0{,}5)(n_{21} + 0{,}5)}$$

nous obtenons un estimateur asymptotiquement sans biais d'ordre $O(n^{-2})$ pour Λ. Donc nous pouvons obtenir de bonnes approximations de Λ. En effet, quand $n \to \infty$, le fait d'ajouter 0,5 a une influence négligeable.

En considérant le logarithme de $\hat{\Lambda}$, nous obtenons

$$\hat{\xi} = \log(\hat{\Lambda}) = \log(n_{11}) - \log(n_{12}) - \log(n_{21}) + \log(n_{22}), \qquad (1.1.3)$$

et nous pouvons, grâce au théorème suivant, déduire ses moments exacts en considérant le vecteur de paramètres inconnus comme étant $\underset{\sim}{\nu}$, voir équation (1.1.1).

Théorème 1.1.2. *Soit $\hat{\xi}$ le logarithme du rapport de cotes estimé (voir équation (1.1.3)) issu d'un tableau de contingence 2×2. Si le vecteur des paramètres inconnus $\underset{\sim}{\nu}$ admet comme distribution une loi de Dirichlet de paramètres $\underset{\sim}{a} = (a_{11}, a_{12}, a_{21}, a_{22})$, alors*

$$\mathbb{E}[\hat{\xi}] = \sum_{(i,j)} c_{ij}\psi(a_{ij}) \quad i,j = 1,2;$$

$$Var[\hat{\xi}] = \sum_{(i,j)} \psi'(a_{ij});$$

où $c_{ij} = 1$ si $i = j = 1,2$, $c_{ij} = -1$ si $i \neq j = 1,2$, $\psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$ et $\psi'(x) = \frac{\partial\psi(x)}{\partial x}$.

Pour une preuve complète de ce théorème, voir Fredette (1999).

Toutefois, l'interprétation du rapport de cotes en tant que mesure d'association peut être difficile. Nous allons voir à la prochaine sous-section la signification du rapport de cotes ainsi qu'une autre mesure d'association plus simple à interpréter qui découle du rapport de cotes.

1.1.5. Le risque relatif

Nous avons affirmé que si $1 < \Lambda < \infty$, les sujets classés dans la première modalité de la variable X ont plus de chance de se retrouver dans la première modalité de la variable Y que ceux de la deuxième modalité de la variable X ; ce qui signifie que $\nu_{1|1} > \nu_{1|2}$. Par exemple quand $\Lambda = 4$, la cote de la première modalité de la variable Y est quatre fois plus élevée pour la première ligne du tableau de contingence que pour la seconde ligne.

Une erreur fréquente est d'interpréter $\Lambda = 4$ comme la probabilité $\nu_{1|1}$ est quatre fois plus élevée que $\nu_{1|2}$. Cette interprétation est valide uniquement si au

lieu de trouver le rapport de cotes, nous calculons le risque relatif (RR).

Définition 1.1.7. *Dans un tableau de contingence* 2×2, *le risque relatif est définit par le ratio*

$$RR = \frac{\nu_{1|1}}{\nu_{1|2}}.$$

Tout comme le rapport de cotes, ce ratio est un nombre non négatif puisque le numérateur et le dénominateur sont non négatifs. Il existe une relation directe entre le rapport de cotes et le risque relatif.

Théorème 1.1.3. $\Lambda = RR\left(\frac{1-\nu_{1|2}}{1-\nu_{1|1}}\right)$

DÉMONSTRATION. Si nous considérons le rapport de cotes des probabilités conditionnelles, on peut écrire

$$\begin{aligned}\Lambda = \frac{\nu_{1|1}\nu_{2|2}}{\nu_{2|1}\nu_{1|2}} &= RR\,\frac{\nu_{2|2}}{\nu_{2|1}} \\ &= RR\left(\frac{1-\nu_{1|2}}{1-\nu_{1|1}}\right).\end{aligned}$$

□

Nous pouvons remarquer que si l'événement correspondant à la première modalité de la variable Y est rare alors les probabilités conditionnelles $\nu_{1|2}$ et $\nu_{1|1}$ seront petites. Dans ce cas, le rapport de cotes est une bonne approximation du risque relatif. En d'autres termes,

$$\lim_{\substack{\nu_{1|2}\to 0 \\ \nu_{1|1}\to 0}} RR \approx \Lambda.$$

Le corollaire suivant énonce une autre propriété du risque relatif liée au rapport de cotes.

Corollaire 1.1.1. *Deux variables aléatoires catégorielles sont indépendantes si et seulement si* $RR = 1$.

DÉMONSTRATION. ($\Longrightarrow$)

Comme les deux variables aléatoires sont indépendantes,

$\nu_{i|j} = \nu_{ji}/\nu_{j+} = (\nu_{j+}\nu_{+i})/\nu_{j+} = \nu_{+i} \quad i = 1, 2$. Le risque relatif devient alors

$$RR = \frac{\nu_{1|1}}{\nu_{1|2}} = \frac{\nu_{+1}}{\nu_{+1}} = 1.$$

($\Longleftarrow$) Nous allons montrer que si $RR = 1$, alors $\Lambda = 1$ et donc les variables sont indépendantes par le théorème 1.1.1. Nous avons

$$\begin{aligned} 1 &= \frac{\nu_{1|1}}{\nu_{1|2}} \\ &\Leftrightarrow \nu_{1|1} = \nu_{1|2} \\ &\Leftrightarrow \frac{\nu_{11}}{\nu_{11} + \nu_{12}} = \frac{\nu_{21}}{\nu_{21} + \nu_{22}} \\ &\Leftrightarrow \nu_{11}\nu_{22} = \nu_{12}\nu_{21} \\ &\Leftrightarrow \Lambda = 1. \end{aligned}$$

□

Afin de calculer une mesure d'association telle le rapport de cotes ou le risque relatif, nous devons avoir accès à des données qui nous permettent de construire le tableau de contingence et ainsi d'estimer les ν_{ij} $i, j = 1, 2$. En ce qui concerne les données qui nous sont disponibles, nous devrons d'abord modéliser les probabilités marginales en ayant recours à la théorie sur les données circulaires.

1.2. DONNÉES CIRCULAIRES

Dans le cadre de notre étude, les données recueillies qui nous seront pertinentes pour modéliser les probabilités marginales ont trait d'une part à l'utilisation du téléphone mobile/cellulaire et d'autre part à la répartition des accidents dans une journée. Dans le premier cas, nous disposons de l'heure du début et de la fin de chaque appel (fait ou reçu) pour chacun des utilisateurs. Chacun de ces deux temps a été quantifié selon le nombre de minutes écoulées depuis minuit. Par exemple, un appel fait à minuit sera associé à la mesure 0, un appel fait à 00h01 aura la valeur 1 et un appel fait à 23h59 sera associé à 1439.

Dans le deuxième cas, nous devrons faire la modélisation de la répartition des accidents dans une journée. Nous utiliserons l'heure des accidents telle qu'inscrite sur le rapport de police. Toutefois comme cette mesure est non précise (voir Baker, 1971), nous avons supposé le fait que l'accident s'est produit dans un intervalle de temps précédant l'heure inscrite. En effet, la mesure fournie par le policier peut être considérée comme une borne supérieure à la véritable mesure car l'accident est constaté au moment où la donnée est inscrite.

Dans cette section, nous justifions la pertinence d'une modélisation de ces données par une approche circulaire. Nous présentons la loi de von Mises qui est le pendant circulaire de la loi normale dans le contexte linéaire. Finalement nous déduirons les estimateurs des deux paramètres impliqués dans une loi de von Mises : μ le paramètre de position et κ le paramètre d'échelle. Les définitions proposées dans cette section sont tirées de Fisher (1993).

1.2.1. Contexte circulaire

Si nous considérons les trois appels cités précédemment (23h59, 00h00, 00h01), nous voyons que l'heure moyenne où les appels ont été faits est minuit. Soit H_i, $i = 1, ..., n$ une variable aléatoire représentant l'heure à laquelle le i^e appel a été fait. Dans un contexte linéaire, nous estimons la moyenne des n observations par la moyenne arithmétique des h_i :

$$\bar{h} = \frac{1}{n}\sum_{i=1}^{n} h_i$$

et la variance par

$$s_h^2 = \frac{1}{n-1}\sum_{i=1}^{n}(h_i - \bar{h})^2.$$

En considérant toujours les mêmes appels, nous avons les mesures suivantes : 0, 1, 1439. Nous obtenons donc $\bar{h} = 480 =$ 08h00 qui n'est pourtant pas l'heure moyenne recherchée de 00h00. Nous obtenons aussi $s_h^2 = 689761$ qui est vraiment trop élevé compte tenu que nos observations sont distantes d'au plus deux minutes. Ces erreurs sont typiquement dues au fait que la variable H considérée en

est une de type circulaire et non de type linéaire.

Définition 1.2.1. *Une variable aléatoire circulaire, ou directionnelle, est une variable aléatoire à valeurs dans l'intervalle $[0, 2\pi)$. Cette variable peut donc être interprétée comme une variable représentant un angle sur un cercle.*

Malgré que nous considérons des mesures de temps qui est un domaine continu, nous avons arrondi les données au nombre de minutes entamées pour effectuer les calculs. Par exemple, pour une valeur exacte de trois minutes 10 secondes, nous avons arrondi la mesure à quatre minutes. Par conséquent la variable $H \in \mathbb{N}$, et elle prend ses valeurs dans l'intervalle discret $[0, 1439]$. Il faut donc convertir chacune des observations selon sa valeur en mesure d'angle. En fait, chaque minute écoulée depuis minuit correspond à une valeur de $\pi/720$ sur un cercle. Nous pouvons considérer une nouvelle variable

$$Y = \frac{\pi H}{720} \in [0, 2\pi).$$

De cette façon, la variable Y correspond à une variable aléatoire circulaire. Toutefois, elle génère des valeurs de $\bar{y} = 2{,}094 = 08h00$ et $s_y = 3{,}624$ qui ne représentent pas encore la réalité.

Dans un contexte circulaire, il est aussi possible de définir des mesures trigonométriques pour résoudre les problèmes de moyenne et de variance échantillonnales soulevées par $\bar{y}$ et s_y^2.

Définition 1.2.2. *Soit $y_1, y_2, ..., y_n$ les valeurs échantillonnales (mesurées sur l'intervalle $[0, 2\pi)$. Le premier moment trigonométrique, noté $\bar{y}_t$, représente la moyenne directionnelle des observations et est définie par :*

$$\bar{y}_t = \begin{cases} \tan^{-1}\left(\frac{S}{C}\right) & si \quad S > 0, C > 0; \\ \tan^{-1}\left(\frac{S}{C}\right) + \pi & si \quad C < 0; \\ \tan^{-1}\left(\frac{S}{C}\right) + 2\pi & si \quad S < 0, C > 0; \end{cases}$$

où $S = \sum_{i=1}^{n} \sin(y_i)$ et $C = \sum_{i=1}^{n} \cos(y_i)$.

Définition 1.2.3. *La variance circulaire échantillonnale, notée V, est définie par $V = 1 - \bar{R}$, $V \in (0, 1)$, où $\bar{R} = R/n$, $R^2 = C^2 + S^2$. Il est à noter que les variables R et $\bar{R}$ représentent respectivement la longueur et la longueur moyenne du vecteur résultant et que $R \in (0, n)$ et $\bar{R} \in (0, 1)$.*
De plus, l'écart type circulaire échantillonale est donné par $v = [-2\log(1 - \bar{R})]^{1/2}$.

Nous avons fourni la figure 1.2.1 qui illustre les quantités $\bar{y}_t$ et R. Sur le graphique, nous avons considéré quatre points sur le cercle représentés par les triangles. En additionnant les vecteurs unitaires associés à chacun des points, nous obtenons la mesure de la direction moyenne ($\bar{y}_t$). La longueur du vecteur reliant le centre du cercle et l'extrémité de la somme des vecteurs unitaires est la longueur résultante (R). La mesure de dispersion V est similaire à la variance dans un contexte linéaire. Plus V est petit, plus la distribution est concentrée. Cependant, il est à noter que lorsque $V = 1$, cela n'implique pas nécessairement que les observations soient réparties uniformément sur le cercle. Un simple exemple suffira à bien justifier cette affirmation.

En effet, si les données recueillies sont les suivantes (en minutes) : 340, 380, 840, 1080, 1320. Après transformation des données en radians, nous obtenons que $R \approx 0$. Par conséquent, $\bar{R} \approx 0$ et $V \approx 1$. Bien que $V \approx 1$, les observations ne sont pas réparties uniformément sur le cercle. Nous pouvons facilement remarquer qu'il y a deux points concentrés autour de 360 minutes et les trois autres points sont concentrés autour de la valeur 1080 minutes. Malgré le fait que $V \approx 1$, V n'est

pas un bon indicateur de la concentration d'un échantillon. Il existe une autre mesure de dispersion qui joue un rôle important dans le calcul d'un intervalle de confiance pour la moyenne directionnelle et pour combiner plusieurs moyennes directionnelles échantillonnales.

Définition 1.2.4. *La dispersion circulaire échantillonnale, notée* $\hat{\delta}$*, est donnée par* $\hat{\delta} = (1 - m_2)/(2\bar{R})$ *où* $m_2 = 1/n \sum_{i=1}^{n} \cos 2(y_i - \bar{y})$ *et* $\bar{R}$ *est tel que défini à la définition 1.2.3.*

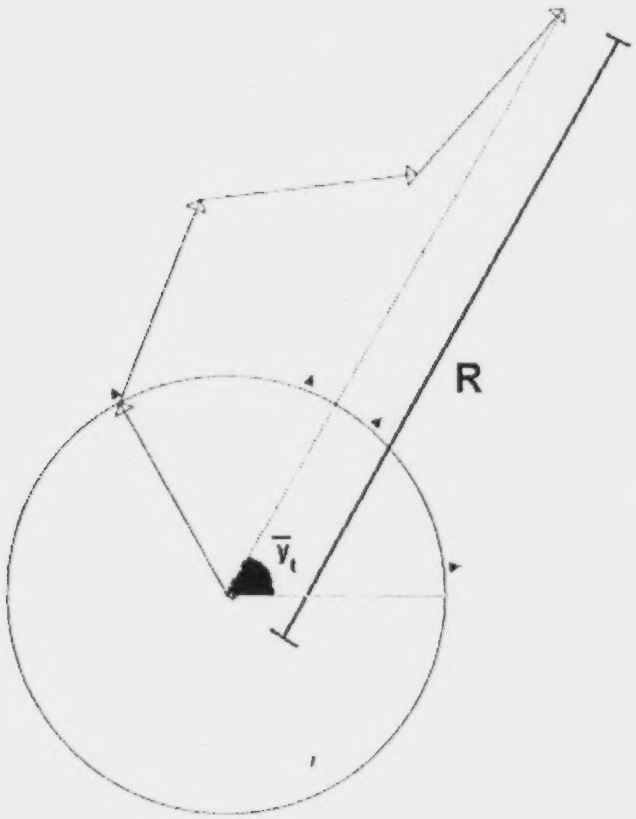

FIGURE 1.2.1. Illustration géométrique de $\bar{y}_t$ et R.

1.2.1.1. *Contexte circulaire paramétrique*

Considérons Y une variable aléatoire continue circulaire définie sur l'intervalle $[0, 2\pi)$. Soient $f(y)$ sa fonction de densité et $F(y)$ sa fonction de répartition. Dans le contexte circulaire, ces deux fonctions possèdent les mêmes propriétés que dans le contexte linéaire, c'est-à-dire :

$$\int_0^{2\pi} f(y)dy = 1 \ ,$$

$$F(y) = P[0 \leq Y \leq y] = \begin{cases} 0 & \text{si} \quad y = 0; \\ \int_0^y f(y)dy & \text{si} \quad y \in (0, 2\pi); \\ 1 & \text{si} \quad y = 2\pi. \end{cases}$$

Aussi, $F(y_2) - F(y_1) = \int_{y_1}^{y_2} f(y)dy$.

Nous pouvons définir la fonction caractéristique de Y pour tout p par

$$\begin{aligned} \phi_p = \mathbb{E}[e^{ipY}] &= \int_0^{2\pi} e^{ipy} f(y)dy \\ &= \int_0^{2\pi} \cos(py) f(y)dy + i \int_0^{2\pi} \sin(py) f(y)dy \\ &= \alpha_p + i\beta_p. \end{aligned}$$

Pour calculer les moments trigonométriques d'une distribution circulaire quelconque, posons

$$\phi_p = \rho_p e^{i\mu_p^0} = \rho_p \cos \mu_p^0 + i\rho_p \sin \mu_p^0 \qquad \forall p.$$

Lorsque $p = 1$, nous écrivons simplement $\rho = \rho_1$ et $\mu_0 = \mu_1^0$. On remarque que μ_0 correspond à la moyenne directionnelle et ρ à la longueur moyenne résultante.

Notons au passage les relations suivantes :

$$\alpha_p = \rho_p \cos \mu_p^0, \qquad \beta_p = \rho_p \sin \mu_p^0,$$

$$\rho_p = \sqrt{\alpha_p^2 + \beta_p^2}, \qquad \mu_p^0 = \tan^{-1}\left(\frac{\alpha_p}{\beta_p}\right).$$

Tout comme dans le contexte non paramétrique, la variance circulaire est définie par $V_0 = 1 - \rho$, $0 \leq V_0 \leq 1$. L'écart type circulaire est donné par $v = \sqrt{-2\log(1 - V_0)}$ et la dispersion circulaire est définie par $\delta = \dfrac{1 - \rho_2}{2\rho^2}$.

1.2.1.2. *Distribution de von Mises*

La distribution de von Mises, notée VM(μ, κ), est une distribution circulaire unimodale et symétrique. C'est le modèle le plus communément utilisé pour des échantillons unimodales provenant de données circulaires. Sa fonction de densité est donnée par :

$$f(y) = \frac{1}{2\pi I_0(\kappa)} \exp[\kappa \cos(y - \mu)]$$

avec $0 \leq y < 2\pi$, $0 \leq \kappa < \infty$ et $I_0(\kappa)$ est la fonction de Bessel modifiée d'ordre 0 définie par :

$$I_0(\kappa) = \sum_{r=0}^{\infty} (r!)^{-2} \left(\frac{\kappa}{2}\right)^{2r}.$$

Les moments trigonométriques d'une variable aléatoire VM(μ, κ) sont donnés par :

moyenne directionnelle : $\mu_0 = \mu$;

longueur moyenne résultante : $\rho = A_1(\kappa)$;

dispersion circulaire : $\delta = [\kappa A_1(\kappa)]^{-1}$;

p^{e} **moment cosinus :** $\alpha_p = A_p(\kappa)$;

p^{e} **moment sinus :** $\beta_p = 0$, $p \geq 1$;

où $A_p(\kappa) = \dfrac{I_p(\kappa)}{I_0(\kappa)}$ et $I_p(\kappa)$ est la fonction de Bessel d'ordre p donnée par

$$I_p(\kappa) = \sum_{r=0}^{\infty} [(r + p)!\, r!]^{-1} \left(\frac{\kappa}{2}\right)^{2r+p}.$$

Notons que, lorsque $\kappa \to 0$, la distribution converge vers une distribution uniforme ; alors que si $\kappa \to \infty$, la distribution tend vers la mesure du point dans la

direction μ. Dans plusieurs cas, la loi de von Mises est l'analogue naturelle sur le cercle de la loi normale sur la droite des réels.

Nous allons maintenant déduire les estimateurs du maximum de vraisemblance pour les paramètres d'une densité de von Mises. Comme nous l'avons vu précédemment, nous devons maximiser la fonction de vraisemblance $L(\mu, \kappa \,|\, \underset{\sim}{y})$. Par définition,

$$\begin{aligned} L(\mu, \kappa \,|\, \underset{\sim}{y}) &= \prod_{i=1}^{n} f(y_i \,|\, \mu, \kappa) \\ &= \frac{1}{(2\pi I_0(\kappa))^n} \exp\left(\kappa \sum_{i=1}^{n} (\cos(y_i - \mu))\right) \\ &= \frac{1}{(2\pi I_0(\kappa))^n} \exp\left(\kappa \left[\cos(\mu) \sum_{i=1}^{n} \cos(y_i) + \sin(\mu) \sum_{i=1}^{n} \sin(y_i)\right]\right) \end{aligned}$$

En prenant le logarithme, nous obtenons

$$\begin{aligned} \log L(\mu, \kappa \,|\, \underset{\sim}{y}) = -n[\log(2\pi) + \log(I_0(\kappa))] \\ + \kappa \left(\cos(\mu) \sum_{i=1}^{n} \cos(y_i) + \sin(\mu) \sum_{i=1}^{n} \sin(y_i)\right) \end{aligned}$$

Pour trouver l'estimateur du maximum de vraisemblance pour le paramètre de position μ, il faut trouver $\hat{\mu}$ qui est solution de

$$\frac{\partial \log L(\mu, \kappa \,|\, \underset{\sim}{y})}{\partial \mu} = 0.$$

Nous avons donc

$$\begin{aligned} \frac{\partial \log L(\mu, \kappa \,|\, \underset{\sim}{y})}{\partial \mu} &= \kappa \left(-\sin(\mu) \sum_{i=1}^{n} \cos(y_i) + \cos(\mu) \sum_{i=1}^{n} \sin(y_i)\right) \\ &= 0 \Longleftrightarrow \frac{\sin(\mu)}{\cos(\mu)} = \frac{\sum_{i=1}^{n} \sin(y_i)}{\sum_{i=1}^{n} \cos(y_i)} \end{aligned}$$

En posant $\mathrm{S} = \sum_{i=1}^{n} \sin(y_i)$ et $\mathrm{C} = \sum_{i=1}^{n} \cos(y_i)$, l'estimateur du maximum de vraisemblance est donné par $\hat{\mu} = \bar{y}_t$ tel que donné à la définition 1.2.2.

Il n'existe pas de forme explicite pour l'estimateur du maximum de vraisemblance de κ, noté $\hat{\kappa}$. Nous pouvons toutefois avoir une bonne approximation de

κ par l'estimation suivante (Fisher, 1993) qui dépend uniquement de $\bar{R}$ tel que donné à la définition 1.2.3.

Proposition 1.2.1. *Une bonne approximation de l'estimateur du maximum de vraisemblance pour κ est donnée par :*

$$\hat{\kappa} = \begin{cases} 2\bar{R} + \bar{R}^3 + \frac{5\bar{R}^5}{6} & si \quad \bar{R} < 0{,}53; \\ \bar{R}^{-3} - 4\bar{R}^2 + 3\bar{R} & si \quad 0{,}53 \leq \bar{R} \leq 0{,}85; \\ -0{,}4 + 1{,}39\bar{R} + \frac{0{,}43}{1-\bar{R}} & si \quad \bar{R} > 0{,}85. \end{cases}$$

1.2.2. Présentation de l'exemple

Nous allons ici présenter un exemple que nous allons utiliser tout au long de ce mémoire afin d'illustrer les méthodes proposées. Nous avons généré, à partir de l'algorithme proposé dans Johnson (1987), 35 données d'une loi de von Mises de paramètre de position $\mu_1 = 1$ et de paramètre de dispersion $\kappa = 0{,}8$ de même que 65 données aussi d'une loi de von Mises mais de paramètre $\mu_2 = 4$ et $\kappa = 0{,}8$. Nous avons placé en annexe A les 100 données. Nous supposerons que ce sont des données représentant l'utilisation du téléphone cellulaire selon le moment de la journée. Nous voyons à la figure 1.2.2 qu'il y a une plus forte utilisation du téléphone cellulaire vers 3h49 et aussi près de 15h17.

Nous pouvons maintenant établir un résumé des mesures paramétriques et non paramétriques qui sont associées à ces données. Nous trouvons

- mesures non paramétriques :

$$S = -6{,}663, \qquad C = -12{,}021,$$

$$\bar{y}_t = 3{,}648, \qquad R = 13{,}744,$$

$$V = 0{,}863, \qquad v = 0{,}544;$$

– mesures paramétriques pour la distribution de von Mises :

$$\mu_0 = 3{,}648, \qquad \hat{k} = 0{,}278,$$

$$\rho = 0{,}137, \qquad \delta = 26{,}216.$$

À la figure 1.2.2, nous retrouvons un graphique comparatif de l'histogramme engendré par ces données et de la densité de von Mises de paramètres de position $\mu_0 = 3{,}648 = 13\text{h}56$ trouvé par la méthode du maximum de vraisemblance et de paramètre de dispersion $\hat{k} = 0{,}278$.

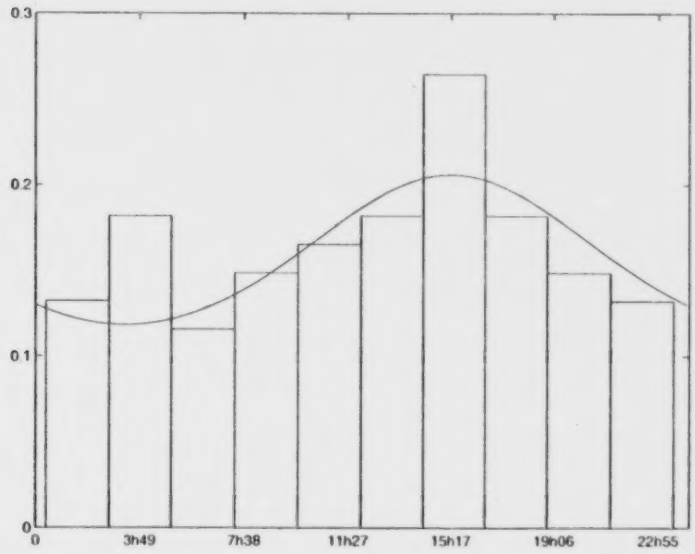

FIGURE 1.2.2. Histogramme comparatif des données de l'exemple et de la densité de VM(3,648 ; 0,278)

Dans ce premier chapitre, nous avons établi les bases statistiques fréquentistes pertinentes en vue de vérifier s'il existe véritablement une association entre deux variables aléatoires catégorielles nominales qualitatives. Nous allons considérer, dans le second chapitre, la théorie bayésienne comme approche plus fidèle pour la modélisation des probabilités nécessaires au calcul éventuel du rapport de cotes et du risque relatif.

Chapitre 2

ESTIMATION DES PARAMÈTRES POUR LA RÉPARTITION DES ACCIDENTS ET DES APPELS SELON LE MOMENT DE LA JOURNÉE

Nous avons vu au premier chapitre qu'il est possible de quantifier le degré d'association entre deux variables aléatoires catégorielles par le rapport de cotes et le risque relatif. Toutefois pour estimer ces mesures, nous devons connaître tous les n_{ij} ou encore tous les $\hat{\nu}_{ij}$, $i, j = 1, 2$. Si nous disposons uniquement d'une seule valeur de n_{ij} en plus de posséder de l'information sur la distribution marginale, alors nous pouvons simuler les autres n_{ij} et ainsi obtenir une estimation de ces mesures d'association.

Dans le présent chapitre, nous présentons d'abord quelques éléments de la théorie bayésienne, principalement issus de Robert (1994, sections 1.2, 1.4 et 3.2) et Leonard et Hsu (1999, section 3.2), qui nous serviront à mieux estimer les paramètres d'une loi de von Mises. Par contre, comme nous remarquons que la distribution que nous voulons estimer n'est pas unimodale, l'ajustement ne donne pas systématiquement des résultats satisfaisants.

À la section suivante, nous proposons comme solution à cette multimodalité, de considérer plutôt un mélange de lois de von Mises afin d'améliorer l'ajustement de la courbe aux données observées. L'algorithme EM (Espérance-Maximisation)

présenté à la section 2.2 sera utilisé pour obtenir l'estimation de chacun des paramètres. Enfin à la section 2.3, nous établirons certains critères pour le choix du modèle à considérer.

2.1. Contexte bayésien

Dans les approches classiques, l'information utilisée pour les analyses provient essentiellement de l'échantillon. Cependant, la théorie bayésienne tente plutôt de combiner l'information fournie par l'échantillon à celle connue avant même de réaliser l'expérience. Nous présentons dans cette section une façon de choisir une densité qui décrira cette information connue *a priori*, comment combiner cette information avec celle contenue dans l'échantillon, et nous présentons aussi une méthode d'estimation des paramètres d'une distribution basée sur la densité marginale, la méthode ML-II.

2.1.1. Éléments de théorie bayésienne

Tout comme au chapitre précédent, nous allons considérer $\underset{\sim}{\theta} \in \Theta$ le vecteur des paramètres inconnus. Dans le but de définir un modèle statistique bayésien, il faut supposer l'existence d'une loi de probabilité *a priori*, notée $\pi(\underset{\sim}{\theta})$ sur les paramètres inconnus.

Définition 2.1.1. *Un modèle statistique bayésien est composé de deux éléments : un modèle statistique paramétrique* $f(\underset{\sim}{y} \mid \underset{\sim}{\theta})$ *et une distribution* a priori $\pi(\underset{\sim}{\theta})$ *sur les paramètres.*

L'ajout de cette loi *a priori* dans le modèle bayésien est ce qui le distingue du modèle statistique classique. En effet, dans certains contextes, il peut arriver que nous ayons déjà de l'information sur un paramètre avant même de récolter les données. Cette connaissance peut avoir été acquise par des études antérieures, par une théorie scientifique ou encore par une opinion subjective. Inclure cette connaissance dans le modèle statistique peut donc être souhaitable.

Avec un modèle bayésien complet, nous pouvons déterminer la loi *a posteriori* qui découle du théorème de Bayes et qui résulte de la combinaison de l'information contenue dans l'échantillon et de celle fournie par la loi *a priori*.

Définition 2.1.2. *Soient $L(\underset{\sim}{\theta} \mid \underset{\sim}{y})$ la fonction de vraisemblance et $\pi(\underset{\sim}{\theta})$ la loi* a priori *sur le vecteur des paramètres. La distribution de $\underset{\sim}{\theta}$ conditionellement au vecteur des observations $\underset{\sim}{y}$, ou la densité* a posteriori, *est donnée par :*

$$\pi(\underset{\sim}{\theta} \mid \underset{\sim}{y}) = \frac{\pi(\underset{\sim}{\theta}) L(\underset{\sim}{\theta} \mid \underset{\sim}{y})}{m(\underset{\sim}{y})},$$

où $m(\underset{\sim}{y}) = \int_{\Theta} \pi(\underset{\sim}{\theta}) L(\underset{\sim}{\theta} \mid \underset{\sim}{y}) \; d\underset{\sim}{\theta}$.

La loi marginale, $m(\underset{\sim}{y})$, joue ici le rôle de constante de normalisation afin que la loi *a posteriori* soit une densité. De plus, un intérêt important de la loi marginale réside principalement dans l'interprétation que nous pouvons en faire. En effet, si $\underset{\sim}{y}$ a comme densité conditionnelle $f(\underset{\sim}{y} \mid \underset{\sim}{\theta})$ et $\underset{\sim}{\theta}$ est une variable aléatoire de densité $\pi(\underset{\sim}{\theta})$ alors $m(\underset{\sim}{y})$ est la densité selon laquelle $\underset{\sim}{y}$ va se produire (voir Berger, 1985, section 3.5).

Au chapitre précédent, nous avons présenté un exemple lié à un contexte circulaire. Nous avons proposé la loi de von Mises pour estimer la densité de nos données. En considérant le même contexte, nous pourrions définir un modèle statistique bayésien complet en spécifiant une loi *a priori* sur les observations et déterminer la densité marginale ainsi que la densité *a posteriori* associées.

Proposition 2.1.1. *Soit $Y_i, \quad i = 1, ..., n$ n variables aléatoires indépendantes de densité $VM(\mu_i, \kappa)$ où κ est connu. Si la densité a priori des μ_i est $VM(\mu_0, \omega\kappa)$, où ω est connu, alors la densité marginale est donnée par*

$$m(\underset{\sim}{y}) = \prod_{i=1}^{n} m(y_i) = \frac{\prod_{i=1}^{n} I_0(\kappa R_i^*)}{(2\pi)^n I_0^n(\kappa) I_0^n(\omega\kappa)},$$

où $R_i^* = \sqrt{C_i^{*2} + S_i^{*2}}$, $C_i^* = \cos y_i + \omega \cos \mu_0$ *et* $S_i^* = \sin y_i + \omega \sin \mu_0$.

Avant de démontrer cette proposition, nous allons examiner les hypothèses et modifier l'écriture de la loi de von Mises prenant le vecteur $\underset{\sim}{y}$ comme données disponibles. Cette réécriture de la densité est suggéré par Mardia et El-Atoum (1976). Nous allons considérer $(\cos y_i, \sin y_i)$ le vecteur aléatoire unitaire associé à y_i prenant ses valeurs sur la surface d'un cercle à deux dimensions S_2 de rayon unitaire et centré à l'origine. Nous pouvons aussi écrire $\underset{\sim}{\mu}_i^* = (\cos \mu_i, \sin \mu_i)$ comme le vecteur de position moyenne. Si

$$f(y_i \mid \underset{\sim}{\mu}_i^*, \kappa) = \begin{cases} \frac{1}{2\pi I_0(\kappa)} \exp\{\kappa \underset{\sim}{\mu}_i^{*t} (\cos y_i, \sin y_i)\} & \text{si} \quad (\cos y_i, \sin y_i) \in S_2; \\ 0 & \text{sinon;} \end{cases}$$

où $\underset{\sim}{\mu}_i^*$ est tel que $\underset{\sim}{\mu}_i^{*t} \underset{\sim}{\mu}_i^* = 1$ et $\kappa > 0$; alors y_i a une densité de von Mises avec $\underset{\sim}{\mu}_i^*$ comme vecteur de direction moyenne et κ comme paramètre de dispersion. Nous pouvons maintenant démontrer la proposition 2.1.1.

DÉMONSTRATION. Par les hypothèses posées à la proposition 2.1.1, nous avons $\underset{\sim}{\mu}_i^* \sim \text{VM}(\underset{\sim}{\mu}_0^* \ , \ \omega\kappa)$, où $\underset{\sim}{\mu}_0^* = (\cos \mu_0, \sin \mu_0)^t$. Nous pouvons maintenant trouver la densité marginale par la relation définie à la définition 2.1.2, c'est-à-dire

$$m(\underset{\sim}{y}) = \prod_{i=1}^{n} m(y_i) = \prod_{i=1}^{n} \int_{\Theta} L(\underset{\sim}{\mu}_i^* \mid y_i, \kappa) \pi(\underset{\sim}{\mu}_i^*) \ d\underset{\sim}{\mu}_i^*.$$

Pour l'observation i, $L(\underset{\sim}{\mu}_i^* | y_i, \kappa) = \frac{1}{2\pi I_0(\kappa)} \exp\{\kappa \underset{\sim}{\mu}_i^{*t} (\cos y_i, \sin y_i)^t\}$ et ainsi nous obtenons

$$\begin{aligned} m(y_i) &= \int_0^{2\pi} \frac{1}{2\pi I_0(\kappa)} \exp\{\kappa \underset{\sim}{\mu}_i^{*t} (\cos y_i, \sin y_i)^t\} \frac{1}{2\pi I_0(\omega\kappa)} \exp\{\omega\kappa \underset{\sim}{\mu}_i^{*t} \underset{\sim}{\mu}_0^*\} \ d\underset{\sim}{\mu}_i^* \\ &= \int_0^{2\pi} \frac{1}{4\pi^2 I_0(\kappa) I_0(\omega\kappa)} \exp\{\kappa \underset{\sim}{\mu}_i^{*t} \left[(\cos y_i, \sin y_i)^t + \omega \underset{\sim}{\mu}_0^*\right]\} \ d\underset{\sim}{\mu}_i^* \\ &= \int_0^{2\pi} \frac{1}{4\pi^2 I_0(\kappa) I_0(\omega\kappa)} \exp\{\kappa (\cos \mu_i, \sin \mu_i) \left[\begin{pmatrix} \cos y_i \\ \sin y_i \end{pmatrix} + \begin{pmatrix} \omega \cos \mu_0 \\ \omega \sin \mu_0 \end{pmatrix}\right]\} \ d\underset{\sim}{\mu}_i^*. \end{aligned}$$

Posons

$$C_i^* = \cos y_i + \omega \cos \mu_0 = R_i^* \cos \tau_i, \tag{2.1.1}$$

$$S_i^* = \sin y_i + \omega \sin \mu_0 = R_i^* \sin \tau_i, \tag{2.1.2}$$

où $R_i^* = \sqrt{C_i^{*2} + S_i^{*2}}$.

À l'aide des équations (2.1.1) et (2.1.2), nous trouvons $\tau_i = \cos^{-1}\left(\frac{C_i^*}{R_i^*}\right)$ et $\tau_i = \sin^{-1}\left(\frac{S_i^*}{R_i^*}\right)$. En divisant l'équation (2.1.2) par l'équation (2.1.1), nous obtenons aussi $\tau_i = \tan^{-1}\left(\frac{S_i^*}{C_i^*}\right)$. Donc nous pouvons écrire,
$\exp\left\{\kappa(\cos\mu_i, \sin\mu_i)\left[\binom{\cos y_i}{\sin y_i} + \binom{\omega\cos\mu_0}{\omega\sin\mu_0}\right]\right\} = \exp\left\{\kappa\underset{\sim}{\mu}_i^{*t}\binom{C_i^*}{S_i^*}\right\}$.
Cependant,

$$\begin{aligned}\underset{\sim}{\mu}_i^{*t}\begin{pmatrix}C_i^*\\S_i^*\end{pmatrix} &= (\cos\mu_i, \sin\mu_i)\begin{pmatrix}R_i^*\cos\tau_i\\R_i^*\sin\tau_i\end{pmatrix}\\ &= R_i^*(\cos\mu_i\cos\tau_i + \sin\mu_i\sin\tau_i)\\ &= R_i^*\cos(\mu_i - \tau_i).\end{aligned}$$

Par conséquent, pour la i^e observation, nous obtenons

$$\begin{aligned}m(y_i) &= \int_0^{2\pi}\frac{1}{4\pi^2 I_0(\kappa)I_0(\omega\kappa)}\exp\{\kappa R_i^*\cos(\mu_i - \tau_i)\}\ d\mu_i\\ &= \frac{I_0(\kappa R_i^*)}{2\pi I_0(\kappa)I_0(\omega\kappa)}\int_0^{2\pi}\frac{1}{2\pi I_0(\kappa R_i^*)}\exp\{\kappa R_i^*\cos(\mu_i - \tau_i)\}\ d\mu_i\\ &= \frac{I_0(\kappa R_i^*)}{(2\pi)I_0(\kappa)I_0(\omega\kappa)},\end{aligned} \tag{2.1.3}$$

car la fonction de la dernière intégrale correspond à une densité $\mathrm{VM}(\tau_i\ ,\ \kappa R_i^*)$.

La densité marginale de $\underset{\sim}{y}$ est finalement donnée par

$$m(\underset{\sim}{y}) = \prod_{i=1}^{n} m(y_i) = \frac{\prod_{i=1}^{n} I_0(\kappa R_i^*)}{(2\pi)^n I_0^n(\kappa)I_0^n(\omega\kappa)}.$$

□

Nous pouvons maintenant trouver la loi *a posteriori* de μ_i pour une observation, sous les mêmes hypothèses, par la relation définie également à la définition 2.1.2. Elle sera donnée par :

$$\pi(\underset{\sim}{\mu}_i^* \mid y_i\ ,\ \kappa\ \omega) = \frac{L(\underset{\sim}{\mu}_i^* \mid y_i\ ,\ \kappa)\pi(\underset{\sim}{\mu}_i^*)}{m(y_i)},$$

où $L(\underset{\sim}{\mu}_i^* \mid y_i\ ,\ \kappa)$ est la fonction de vraisemblance des observations donnée par $f(y_i \mid \underset{\sim}{\mu}_i^*\ ,\ \kappa)$ et $m(y_i)$ est la densité marginale donnée à l'équation (2.1.3). Ainsi

nous obtenons,

$$\pi(\mu_i^* \mid y_i\ ,\ \kappa\,\omega) = \frac{\frac{1}{4\pi^2 I_0(\kappa) I_0(\omega\kappa)} \exp\{\kappa R_i^* \cos(\mu_i - \tau_i)\}}{\frac{I_0(\kappa R_i^*)}{2\pi I_0(\kappa) I_0(\omega\kappa)}}$$

$$= \frac{1}{2\pi I_0(\kappa R_i^*)} \exp\{\kappa R_i^* \cos(\mu_i - \tau_i)\}, \qquad (2.1.4)$$

qui correspond à une densité de von Mises de paramètres τ_i et κR_i^*.

Nous avons considéré certaines hypothèses pour déduire les lois marginale et *a posteriori* impliquées dans le modèle statistique bayésien. Pourtant, une des difficultés reliées à la théorie bayésienne est de choisir une loi *a priori* adéquate avec une bonne paramétrisation. Il existe des techniques afin de pallier à ce problème dont l'utilisation d'une loi *a priori* provenant d'une famille conjuguée de lois.

Définition 2.1.3. *Une famille $\mathcal{F}$ de lois sur Θ est dite conjuguée si, pour tout $\pi \in \mathcal{F}$, la loi* a posteriori $\pi(\underset{\sim}{\theta} \mid \underset{\sim}{y})$ *appartient également à $\mathcal{F}$.*

En somme, si une loi *a priori* provient d'une famille conjuguée de lois, alors le passage de cette loi à la loi *a posteriori* résulte uniquement en un changement de paramètres. Un des avantages importants de ce choix tient du fait que nous pouvons toujours calculer la loi *a posteriori*. Raiffa et Schlaifer (1961), justifient la pertinence de cette approche par un raisonnement d'invariance. En effet, quand les observations $\underset{\sim}{y} \sim f(\underset{\sim}{y}|\underset{\sim}{\theta})$ modifient $\pi(\underset{\sim}{\theta})$ à $\pi(\underset{\sim}{\theta}|\underset{\sim}{y})$, l'information apportée par le vecteur $\underset{\sim}{y}$ sur θ est assurément limitée. En choisissant une loi *a priori* conjuguée, nous ne remettons pas en question la pertinence de la loi π mais plutôt sa paramétrisation.

On remarque qu'à l'exemple précédent, si la densité d'une observation est $\mathrm{VM}(\mu_i, \kappa)$ et si la loi *a priori* de μ_i est $\mathrm{VM}(\mu_0, \omega\kappa)$ alors la loi *a posteriori* est donnée par

$$\frac{1}{2\pi I_0(\kappa R_i^*)} \exp\{\kappa R_i^* \cos(\mu_i - \tau_i)\}$$

qui est $VM(\tau_i, \kappa R_i^*)$. Donc en prenant une loi *a priori* de von Mises conjuguée plutôt que de modifier la forme de la densité elle-même, nous ajustons uniquement les paramètres de cette loi.

2.1.2. Description de la méthode ML-II

Quand l'espace paramètre Θ est non dénombrable, comme par exemple un intervalle, le choix de la loi *a priori* devient plus compliqué. Plusieurs méthodes ont été développées pour construire une loi *a priori* basée sur les observations. Nous allons ici utiliser la densité marginale $m(\underset{\sim}{y})$ dans la construction de cette loi *a priori*. Cette méthode consiste essentiellement en l'estimation des paramètres de la loi *a priori* par la maximisation du logarithme de la loi marginale. Cette méthode de construction de $\pi(\underset{\sim}{\theta})$ produit ce qui est appelée communément la loi *a priori* du maximum de vraisemblance de type II (ML-II).

Définition 2.1.4. *Soit Γ une classe de lois a priori à considérer. Si $\hat{\pi} \in \Gamma$ satisfait (pour le vecteur des données observées $\underset{\sim}{y}$)*

$$m(\underset{\sim}{y} \mid \hat{\pi}) = \sup_{\pi \in \Gamma} m(\underset{\sim}{y} \mid \pi),$$

alors $\hat{\pi}$ est appelé loi a priori du maximum de vraisemblance de type II ou loi a priori ML-II.

Si nous considérons encore les mêmes hypothèses que précédemment, nous avons montré que la loi marginale pour une observation est donnée par

$$m(y_i) = \frac{I_0(\kappa R_i^*)}{2\pi I_0(\kappa) I_0(\omega\kappa)}.$$

Nous nous intéressons maintenant à maximiser cette densité marginale $m(\underset{\sim}{y})$ par rapport à tous les paramètres impliqués ; c'est-à-dire μ_0, ω et κ. Nous allons maximiser ces paramètres par la méthode du maximum de vraisemblance. Il nous faut alors maximiser le logarithme de la fonction de vraisemblance de cette loi marginale.

Proposition 2.1.2. *Soit Y_i une variable aléatoire indépendante pour $i = 1, \ldots, n$ de densité $VM(\mu_i, \kappa)$. Si $\mu_i \sim VM(\mu_0, \omega\kappa)$ alors les estimateurs du maximum de vraisemblance des paramètres μ_0, ω et κ obtenus par la méthode ML-II sont respectivement donnés par :*

- *$\hat{\mu_0}$ solution de $\sum_{i=1}^{n} \frac{A_1(\kappa R_i^*)}{\sqrt{R_i^*}} \sin(y_i - \mu_0) = 0$;*

- *$\hat{\omega}$ solution de $\frac{1}{n} \sum_{i=1}^{n} \left(\frac{A_1(\kappa R_i^*)[\omega + \cos(y_i - \mu_0)]}{\sqrt{R_i^*}} \right) = A_1(\omega\kappa)$;*

- *$\hat{\kappa}$ solution de $\frac{1}{n} \sum_{i=1}^{n} R_i^* A_1(\kappa R_i^*) = A_1(\kappa) + A_1(\omega\kappa)$;*

où $R_i^* = \sqrt{C_i^{*2} + S_i^{*2}}$, $C_i^* = \cos y_i + \omega \cos \mu_0$, $S_i^* = \sin y_i + \omega \sin \mu_0$ et $A_1(\cdot) = \frac{I_1(\cdot)}{I_0(\cdot)}$.

DÉMONSTRATION. Le logarithme de la loi marginale est donné par :

$$\begin{aligned} \log L(\mu_0 \, , \, \omega \, , \, \kappa \mid \underset{\sim}{y}) &= \log \prod_{i=1}^{n} m(y_i) \\ &= \sum_{i=1}^{n} \log[I_0(\kappa R_i^*)] - n \log[2\pi] - n \log[I_0(\kappa)] - n \log[I_0(\omega\kappa)]. \end{aligned}$$

Si nous nous intéressons d'abord au paramètre μ_0, il faut trouver $\hat{\mu_0}$ tel que

$$\begin{aligned} \frac{\partial \log L(\mu_0 \, , \, \omega \, , \, \kappa \mid \underset{\sim}{y})}{\partial \mu_0} &= 0 \\ &\Leftrightarrow \kappa \sum_{i=1}^{n} A_1(\kappa R_i^*) \frac{\partial R_i^*}{\partial \mu_0} = 0. \end{aligned}$$

Or, nous avons

$$\begin{aligned} \frac{\partial R_i^*}{\partial \mu_0} &= \frac{1}{\sqrt{R_i^*}} \omega (S_i^* \cos \mu_0 - C_i^* \sin \mu_0) \\ &= \frac{\omega}{\sqrt{R_i^*}} (\cos \mu_0 \sin y_i - \sin \mu_0 \cos y_i) \\ &= \frac{\omega}{\sqrt{R_i^*}} \sin(y_i - \mu_0). \end{aligned}$$

Donc, nous obtenons

$$\frac{\partial \log L(\mu_0\ ,\ \omega\ ,\ \kappa \mid y)}{\partial \mu_0} = 0 \Leftrightarrow \sum_{i=1}^{n} \frac{A_1(\kappa R_i^*)}{\sqrt{R_i^*}} \sin(y_i - \mu_0) = 0.$$

Si nous nous intéressons plutôt au paramètre ω, nous devons trouver $\hat{\omega}$ solution de

$$\frac{\partial \log L(\mu_0\ ,\ \omega\ ,\ \kappa \mid y)}{\partial \omega} = 0$$
$$\Leftrightarrow \kappa \sum_{i=1}^{n} A_1(\kappa R_i^*) \frac{\partial R_i^*}{\partial \omega} - n\kappa A_1(\omega\kappa) = 0.$$

Or, nous avons

$$\begin{aligned}
\frac{\partial R_i^*}{\partial \omega} &= \frac{1}{\sqrt{R_i^*}}(C_i^* \cos \mu_0 + S_i^* \sin \mu_0) \\
&= \frac{1}{\sqrt{R_i^*}}(\omega + \cos y_i \cos \mu_0 + \sin y_i \sin \mu_0) \\
&= \frac{1}{\sqrt{R_i^*}}[\omega + \cos(y_i - \mu_0)].
\end{aligned}$$

Donc, nous obtenons

$$\frac{\partial \log L(\mu_0\ ,\ \omega\ ,\ \kappa \mid y)}{\partial \omega} = 0 \Leftrightarrow \frac{1}{n} \sum_{i=1}^{n} \frac{A_1(\kappa R_i^*)}{\sqrt{R_i^*}} [\omega + \cos(y_i - \mu_0)] = A_1(\omega\kappa).$$

Finalement, pour le paramètre κ, nous devons trouver $\hat{\kappa}$ tel que

$$\frac{\partial \log L(\mu_0\ ,\ \omega\ ,\ \kappa \mid y)}{\partial \kappa} = 0$$
$$\Leftrightarrow \sum_{i=1}^{n} [R_i^* A_1(\kappa R_i^*)] - n[A_1(\kappa) + \omega A_1(\omega\kappa) = 0;$$
$$\Leftrightarrow \frac{1}{n} \sum_{i=1}^{n} R_i^* A_1(\kappa R_i^*) = A_1(\kappa) + \omega A_1(\omega\kappa).$$

□

Nous avons maintenant les outils nécessaires afin d'estimer les valeurs des paramètres impliqués dans un modèle statistique bayésien où chacune des observations provient d'une loi de VM(μ_i , κ) avec une loi *a priori* VM($\mu_0, \omega\kappa$) sur les paramètres de position μ_i.

Si nous examinons la figure 1.2.2 des données de l'exemple, nous remarquons que l'allure générale de l'histogramme semble plutôt multimodale. L'ajustement de la courbe par la méthode ML-II pour cet exemple n'est pas si mal. Toutefois, pour pallier à ce manque d'ajustement, nous proposons dans la section suivante de considérer un modèle de mélange de loi de von Mises.

2.2. Mélange de lois

Quand nous nous retrouvons avec un histogramme des données qui présente plusieurs modes distincts, une solution est de considérer un mélange de lois unimodales. Dans cette section, nous débuterons par définir ce qu'est un mélange de lois. Nous présenterons l'algorithme EM (Espérance-Maximisation) qui sera utilisé pour estimer chacun des paramètres impliqués dans le mélange. Par la suite, nous déterminerons un critère qui nous permettra de choisir le nombre de composantes à conserver pour le modèle en fonction du nombre de paramètres à estimer.

De façon générale, un mélange de lois à g composantes pour une variable aléatoire Y s'écrit sous la forme :

$$f(y|\underset{\sim}{\psi}_g) = \sum_{j=1}^{g} p_j f_j(y \mid \underset{\sim}{\theta}_j),$$

avec $\underset{\sim}{\theta}_j$ le vecteur des paramètres de la loi considérée ; $p_j > 0$, $j = 1, \ldots, m$ les probabilités associées à chacune des lois satisfaisant la contrainte $\sum_{j=1}^{g} p_j = 1$. Le vecteur de tous les paramètres s'écrit $\underset{\sim}{\psi}_g = (p_1, \ldots, p_g, \underset{\sim}{\theta}_1, \ldots, \underset{\sim}{\theta}_g)$ et finalement $f_j(y \mid \underset{\sim}{\theta}_j) \geq 0$ et $\int_0^{2\pi} f_j(y \mid \underset{\sim}{\theta}_j)\, dy = 1$, pour $j = 1, \ldots, g$.

Nous avons montré à la proposition 2.1.1 que si $Y_i, \quad i = 1, \ldots, n$ sont n variables aléatoires indépendantes de densité $\mathrm{VM}(\mu_i\ ,\ \kappa)$ et si la densité *a priori* des μ_i est $\mathrm{VM}(\mu_0\ ,\ \omega\kappa)$, alors nous avons

$$m(\underset{\sim}{y}) = \frac{\prod_{i=1}^{n} I_0(\kappa R_i^*)}{(2\pi)^n I_0^n(\kappa) I_0^n(\omega\kappa)}.$$

Dans le but d'appliquer encore une fois la méthode ML-II pour l'estimation des paramètres, nous allons nous intéresser à un mélange de lois marginales de cette forme. Dans ce cas, nous avons $\underset{\sim}{\theta}_j = (\mu_j, \omega_j, \kappa)$ et $f_j(\underset{\sim}{y} \mid \underset{\sim}{\theta}_j) = m_j(\underset{\sim}{y})$.

Pour ajuster le modèle, nous considérons g densités marginales données par l'équation (2.1.3) pour lesquelles nous devons estimer le vecteur des paramètres inconnus $\underset{\sim}{\psi}_g = (p_1, \ldots, p_g, \underset{\sim}{\theta}_1, \ldots, \underset{\sim}{\theta}_g)$ par $\hat{\underset{\sim}{\psi}}_g = (\hat{p}_1, \ldots, \hat{p}_g, \hat{\underset{\sim}{\theta}}_1, \ldots, \hat{\underset{\sim}{\theta}}_g)$. Pour ce faire, nous allons utiliser l'algorithme EM (Espérance-Maximisation) fréquemment utilisé dans la théorie des mélanges de lois. Toutefois, nous allons d'abord présenter des résultats sur l'identifiabilité des paramètres dans un tel contexte.

2.2.1. Identifiabilité

Une des questions fondamentales concernant les mélanges de certaines lois est l'identifiabilité des paramètres. Gumbel (1954), Jones et James (1969), Mardia et Sutton (1975) comptent parmi ceux qui suggèrent l'utilisation de modèles de mélanges de lois reliées au contexte circulaire pour ajuster une courbe qui présente une allure multimodale. Cette propriété d'identifiabilité est une considération importante pour l'estimation des paramètres d'un mélange de lois (Maritz, 1970). Tout d'abord, définissons ce que nous entendons par identifiabilité (voir Yakowitz et Spragins (1968), Théorème A).

Définition 2.2.1. *Soit* $\{f(y \mid \underset{\sim}{\theta}) : \underset{\sim}{\theta} \in \Theta\}$ *une classe de fonctions de densité continues, où* Θ *est l'espace des paramètres. Cette classe est identifiable si pour tout ensemble distinct de paramètres* $\underset{\sim}{\theta}_1, \ldots, \underset{\sim}{\theta}_l$, $l \geq 1$ *et de réels* $p_1, \ldots, p_l$, *la relation* $\sum_{i=1}^{l} p_i f(y \mid \underset{\sim}{\theta}_i) \equiv 0$ *implique que* $p_i = 0, \ \forall i$.

Théorème 2.2.1. *La classe des distributions de von Mises est identifiable.*

Pour une preuve complète de ce théorème, voir Fraser, Hsu et Walker (1981).

Comme nous savons maintenant que la classe des distributions de von Mises est identifiable, nous allons appliquer l'algorithme EM dans le but d'estimer chacun des paramètres présents dans le mélange de lois.

2.2.2. Algorithme EM

Comme nous avons vu précédemment à la section 1.2.1 du premier chapitre, nous pouvons tenter d'estimer chacun des paramètres du vecteur $\underset{\sim}{\theta}$ par la méthode du maximum de vraisemblance. Dans le cas d'un mélange de lois, il nous faut maximiser le logarithme de la fonction de vraisemblance de chacune des densités marginales pour chacun des paramètres, c'est-à-dire :

$$\max_{\underset{\sim}{\psi}_g} l(\underset{\sim}{\psi}_g \mid \underset{\sim}{y}) = \max_{\underset{\sim}{\psi}_g} \log L(\underset{\sim}{\psi}_g \mid \underset{\sim}{y}) = \max_{\underset{\sim}{\psi}_g} \sum_{i=1}^{n} \left(\log \sum_{j=1}^{g} p_j f_j(y_i \mid \underset{\sim}{\theta}_j) \right).$$

La valeur de $\hat{\underset{\sim}{\psi}}_g$ qui maximise $l(\underset{\sim}{\psi}_g \mid \underset{\sim}{y})$ est solution de

$$\frac{\partial l(\underset{\sim}{\psi}_g \mid \underset{\sim}{y})}{\partial \underset{\sim}{\psi}_g} = 0,$$

où $\partial \underset{\sim}{\psi}_g$ signifie que nous dérivons $l(\underset{\sim}{\psi}_g \mid \underset{\sim}{y})$ pour chacun des paramètres inclus dans $\underset{\sim}{\psi}_g$ et ensuite résoudre simultanément le système d'équations qui est ainsi formé.

Dans le contexte d'un mélange de lois marginales tel que celui auquel nous sommes confrontés, les équations trouvées à la proposition 2.1.2 formant le système à résoudre sont non linéaires en $\underset{\sim}{\psi}_g$, ce qui implique qu'aucune forme explicite ne peut être déduite pour les estimateurs. Toutefois, dans le but de trouver les estimateurs, nous pouvons utiliser une méthode itérative : l'algorithme EM (Espérance-Maximisation) qui a été introduit par Dempster, Laird et Rubin (1977) et décrit pour les mélanges de lois dans Titterington, Smith et Makov

(1985, section 4.3).

2.2.2.1. *Modèle augmenté*

Toujours dans la situation où nous sommes intéressés à un mélange de densités marginales tel que défini précédemment, nous pouvons introduire des variables latentes $\underset{\sim}{z}_i = (z_{i1}, \ldots, z_{ig})^t$. Ce vecteur de longueur g possède un 1 à la position j correspondant à la densité $f_j(y_i \mid \underset{\sim}{\theta}_j)$ à laquelle appartient l'observation i et des 0 partout ailleurs. Nous pouvons donc considérer chaque $\underset{\sim}{z}_i$ comme étant une variable multinomiale (avec $N = 1$) nous indiquant de quelle loi provient la i^e observation. Les données complètes sont alors de la forme

$$\{\underset{\sim}{x}_i, \quad i = 1, \ldots, n\} = \{(y_i\ ,\ \underset{\sim}{z}_i), \quad i = 1, \ldots, n\}.$$

Il est à noter que la variable $\underset{\sim}{x}_i$ est indépendante de $\underset{\sim}{x}_j$ pour tout $i \neq j$. De la même façon, la variable $\underset{\sim}{z}_k$ est indépendante de $\underset{\sim}{z}_l \quad \forall k \neq l$.

Pour commencer, nous pouvons écrire la fonction de vraisemblance sous la forme :

$$L(\underset{\sim}{\psi}_g \mid \underset{\sim}{x}) = \prod_{i=1}^{n} \prod_{j=1}^{g} p_j^{z_{ij}} f_j^{z_{ij}}(y_i \mid \underset{\sim}{\theta}_j).$$

En prenant le logarithme, nous devons maximiser la fonction

$$\begin{aligned} l(\underset{\sim}{\psi}_g \mid \underset{\sim}{x}\ ) &= \sum_{i=1}^{n} \sum_{j=1}^{g} (z_{ij}\ \log\ p_j \quad + \quad z_{ij}\ \log f_j(y_i \mid \underset{\sim}{\theta}_j)) \\ &= \sum_{i=1}^{n} \underset{\sim}{z}_i^t\, V(\underset{\sim}{p}) \quad + \quad \sum_{i=1}^{n} \underset{\sim}{z}_i^t\, U_i(\underset{\sim}{\theta}), \end{aligned}$$

où $V(\underset{\sim}{p})$ a comme j^e composante l'élément $\log\ p_j$ et $U_i(\underset{\sim}{\theta})$ a comme j^e composante l'élément $\log f_j(y_i \mid \underset{\sim}{\theta}_j)$.

Pour arriver à trouver $\hat{\underset{\sim}{\psi}}_g$ qui maximise $l(\underset{\sim}{\psi}_g \mid \underset{\sim}{x})$, nous devons connaître implicitement $\underset{\sim}{z}$. Comme ce sont des variables latentes, elles nous sont donc inconnues. Nous parviendrons toutefois à trouver $\hat{\underset{\sim}{\psi}}_g$ par itération à l'aide de l'algorithme EM.

2.2.2.2. *Étapes de l'algorithme EM*

Cet algorithme consiste à générer une suite d'estimateurs $\{\underset{\sim}{\psi}_g^{(k)}\}$ à partir d'un estimateur initial $\{\underset{\sim}{\psi}_g^{(0)}\}$ jusqu'à obtenir la convergence vers $\hat{\underset{\sim}{\psi}}_g$. Chaque itération consiste en deux étapes :

(1) Étape E (Espérance)

$$\text{Calculer} \quad Q(\underset{\sim}{\psi}_g\ ,\ \underset{\sim}{\psi}_g^{(k)}) = \mathbb{E}^{\underset{\sim}{z}\ |\ \underset{\sim}{\psi}_g^{(k)}\ ,\ \underset{\sim}{y}}[l(\underset{\sim}{\psi}_g \mid \underset{\sim}{y}\ ,\ \underset{\sim}{z})].$$

Nous obtenons la distribution de $\underset{\sim}{z} \mid \underset{\sim}{\psi}_g^{(k)}\ ,\ \underset{\sim}{y}$ par la relation suivante :

$$f(\underset{\sim}{z} \mid \underset{\sim}{\psi}_g^{(k)}\ ,\ \underset{\sim}{y}) = \frac{f(\underset{\sim}{z}\ ,\ \underset{\sim}{y} \mid \underset{\sim}{\psi}_g^{(k)})}{f(\underset{\sim}{y} \mid \underset{\sim}{\psi}_g^{(k)})} \propto f(\underset{\sim}{y}\ ,\ \underset{\sim}{z} \mid \underset{\sim}{\psi}_g^{(k)}).$$

De plus, nous utilisons le fait que

$$\begin{aligned} l(\underset{\sim}{\psi}_g \mid \underset{\sim}{y}\ ,\ \underset{\sim}{z}) &= \log L(\underset{\sim}{\psi}_g \mid \underset{\sim}{y}\ ,\ \underset{\sim}{z}) = \log f(\underset{\sim}{y}\ ,\ \underset{\sim}{z} \mid \underset{\sim}{\psi}_g) \\ &= \log f(\underset{\sim}{y} \mid \underset{\sim}{z}\ ,\ \underset{\sim}{\psi}_g) f(\underset{\sim}{z} \mid \underset{\sim}{\psi}_g). \end{aligned}$$

Donc, l'étape E devient :

$$\text{calculer} \quad Q(\underset{\sim}{\psi}_g\ ,\ \underset{\sim}{\psi}_g^{(k)}) = \mathbb{E}^{\underset{\sim}{z}\ |\ \underset{\sim}{\psi}_g^{(k)}\ ,\ \underset{\sim}{y}}[\log[f(\underset{\sim}{y} \mid \underset{\sim}{z},\ \underset{\sim}{\psi}_g) f(\underset{\sim}{Z} \mid \underset{\sim}{\psi}_g)]].$$

Il nous faut cependant trouver les densités de $\underset{\sim}{y} \mid \underset{\sim}{z}\ ,\ \underset{\sim}{\psi}_g$ et $\underset{\sim}{z} \mid \underset{\sim}{\psi}_g$.

(2) Étape M (Maximisation)

Trouver $\underset{\sim}{\psi}_g = \underset{\sim}{\psi}_g^{(k+1)}$ qui maximise $Q(\underset{\sim}{\psi}_g\ ,\ \underset{\sim}{\psi}_g^{(k)})$ en résolvant

$$\frac{\partial Q(\underset{\sim}{\psi}_g\ ,\ \underset{\sim}{\psi}_g^{(k)})}{\partial \underset{\sim}{\psi}_g} = 0.$$

2.2.2.3. *Algorithme EM pour un mélange de densités marginales à g composantes*

Dans le but d'appliquer l'algorithme EM à un mélange de g lois marginales, nous considérons la densité suivante :

$$f(y_i \mid \underset{\sim}{\psi}_g) = f(y_i \mid \underset{\sim}{\theta}_1 , \underset{\sim}{\theta}_2 , \ldots , \underset{\sim}{\theta}_g , p_1 , p_2, \ldots , p_g)$$
$$= p_1 f_1(y_i \mid \underset{\sim}{\theta}_1) + p_2 f_2(y_i \mid \underset{\sim}{\theta}_2) + \ldots + (1 - \sum_{j=1}^{g-1} p_j) f_g(y_i \mid \underset{\sim}{\theta}_g).$$

Après avoir introduit les variables latentes $\underset{\sim}{z}_i$, nous obtenons le modèle augmenté

$$f(y_i \mid \underset{\sim}{z}_i , \underset{\sim}{\psi}_g) = \left[\prod_{j=1}^{g-1} f_j(y_i \mid \theta_j)^{z_{ij}}\right] f_g(y_i \mid \theta_g)^{1-\sum_{l=1}^{g-1} z_{il}}$$

où $\underset{\sim}{z}_i = \underset{\sim}{z}_i \mid \underset{\sim}{\psi}_g = \underset{\sim}{z}_i \mid p_1 , p_2 , \ldots , p_{g-1}$ tel que

$$\mathbb{P}(\underset{\sim}{z}_i \mid p_1 , p_2 , \ldots , p_{g-1}) = \left[\prod_{j=1}^{g-1} p_j^{z_{ij}}\right] \left(1 - \sum_{l=1}^{g-1} p_l\right)^{1-\sum_{l=1}^{g-1} z_{il}} , \quad i = 1, \ldots , n.$$

Puisque chaque $\underset{\sim}{z}_i$ est en fait une variable multinomiale (avec N=1) nous indiquant de quelle loi provient cette i^{e} observation, si nous considérons $f_j(\cdot)$ comme étant un succès alors pour un p_j fixé, $z_{ij} \mid p_j \sim Bin(1 , p_j)$, $j = 1, \ldots , g$. Ainsi, nous remarquons que z_{i1} prendra la valeur 1 dans une proportion p_1 et alors nous avons $f(y_i \mid z_{i1} , \underset{\sim}{\psi}_g) = f_1(y_i \mid z_{i1} , \underset{\sim}{\theta}_1)$. Nous pouvons donc dire de façon générale que z_{ij} prendra la valeur 1 dans une proportion p_j et alors $f(y_i \mid z_{ij} , \underset{\sim}{\psi}_g) = f_j(y_i \mid z_{ij} , \underset{\sim}{\theta}_j)$. Nous constatons que $y_i \mid z_{i1} , \underset{\sim}{\psi}_m$ ne dépend pas des p_j, $j = 1 , \ldots , g$. De plus, la densité de $\underset{\sim}{z}_i \mid \underset{\sim}{\psi}_g$ dépend uniquement des probabilités p_j, $j = 1, \ldots , g-1$ et non des $\underset{\sim}{\theta}_j$, $j = 1 , \ldots , g$.

Afin de pouvoir appliquer l'algorithme EM, nous devons trouver la densité de $\underset{\sim}{z} \mid \underset{\sim}{\psi}_g^{(k)} , \underset{\sim}{y}$ nécessaire à l'étape E (Espérance).

Théorème 2.2.2. *Dans un mélange de lois à g composantes, la loi de* $\underset{\sim}{z}_i \mid \underset{\sim}{\psi}_g^{(k)} , \underset{\sim}{y}$ *est donnée par*

$$\underset{\sim}{z}_i \mid \underset{\sim}{\psi}_g^{(k)} , \underset{\sim}{y} \quad = \underset{\sim}{z}_i \mid \underset{\sim}{\theta} , \underset{\sim}{p}_{(g-1)} , \underset{\sim}{y} \sim \mathcal{M}\left(g, \frac{p_j f_j(y_i \mid \underset{\sim}{\theta}_j)}{\sum_{r=1}^{g} p_r f_r(y_i \mid \underset{\sim}{\theta}_r)}\right)$$

où $f_j(y_i \mid \underset{\sim}{\theta}_j)$ est la densité de la j^e composante, $\underset{\sim}{\theta} = (\underset{\sim}{\theta}_1, \ldots, \underset{\sim}{\theta}_g)$ et $\underset{\sim}{p}_{(g-1)} = (p_1, \ldots, p_{g-1})$.

DÉMONSTRATION. Supposons que nous avons un modèle à g composantes :

$$\begin{aligned} f(\underset{\sim}{z}_i \mid \underset{\sim}{\theta},\ \underset{\sim}{p}_{(g-1)},\ \underset{\sim}{y}) &= f(\underset{\sim}{z} \mid y_i,\ \underset{\sim}{\theta},\ \underset{\sim}{p}_{(g-1)}) \\ &= \frac{f(y_i \mid \underset{\sim}{z}_i,\ \underset{\sim}{\theta},\ \underset{\sim}{p}_{(g-1)}) f(\underset{\sim}{z}_i \mid \underset{\sim}{\theta},\ \underset{\sim}{p}_{(g-1)})}{f(y_i \mid \underset{\sim}{\theta},\ \underset{\sim}{p}_{(g-1)})} \\ &\propto f(y_i \mid \underset{\sim}{z}_i,\ \underset{\sim}{\theta}) f(\underset{\sim}{z}_i \mid \underset{\sim}{p}_{(g-1)}) \\ &= [p_1 f_1(y_i \mid \underset{\sim}{\theta}_1)]^{z_{i1}} \cdots [(1 - \sum_{j=1}^{g-1} p_j) f_g(y_i \mid \underset{\sim}{\theta}_g)]^{(1-\sum_{j=1}^{g-1} z_{ij})} \\ &\propto \prod_{j=1}^{g} \left(\frac{p_j f_j(y_i \mid \underset{\sim}{\theta}_j)}{\sum_{r=1}^{g} p_r f_r(y_i \mid \underset{\sim}{\theta}_r)} \right)^{z_{ij}}, \end{aligned}$$

où $p_g = 1 - \sum_{j=1}^{g-1} p_j$ et $\sum_{j=1}^{g} z_{ij} = 1$.
Par conséquent, nous obtenons

$$\underset{\sim}{z}_i \mid \underset{\sim}{\theta},\ \underset{\sim}{p}_{(g-1)},\ \underset{\sim}{y} \sim \mathcal{M}\left(g, \frac{p_j f_j(y_i \mid \underset{\sim}{\theta}_j)}{\sum_{r=1}^{g} p_r f_r(y_i \mid \underset{\sim}{\theta}_r)} \right).$$

□

Nous pouvons maintenant effectuer l'étape E (Espérance) :

Proposition 2.2.1.

$$Q(\underset{\sim}{\psi}_g\ ,\ \underset{\sim}{\psi}_g^{(k)}) = \sum_{j=1}^{g} \sum_{i=1}^{n} \mathbb{E}^{\underset{\sim}{z} \mid \underset{\sim}{\psi}_g^{(k)},\ \underset{\sim}{y}}[Z_{ij}] \log f_j(y_i \mid \underset{\sim}{\theta}_j) + \sum_{j=1}^{g} \log p_j \sum_{i=1}^{n} \mathbb{E}^{\underset{\sim}{z} \mid \underset{\sim}{\psi}_g^{(k)},\ \underset{\sim}{y}}[Z_{ij}],$$

où $\mathbb{E}^{\underset{\sim}{z} \mid \underset{\sim}{\psi}_g^{(k)},\ \underset{\sim}{y}}[Z_{ig}] = 1 - \sum_{j=11}^{g-1} \mathbb{E}^{\underset{\sim}{z} \mid \underset{\sim}{\psi}_g^{(k)},\ \underset{\sim}{y}}[Z_{ij}]$ et $p_g = \sum_{j=1}^{g-1} p_j$.

DÉMONSTRATION. Par définition, nous avons

$$
\begin{aligned}
Q(\underset{\sim}{\psi}_g\ ,\ \underset{\sim}{\psi}_g^{(k)}) &= \mathbb{E}^{\underset{\sim}{z}\ |\ \underset{\sim}{\psi}_g^{(k)}\ ,\ \underset{\sim}{y}}[\log(f(\underset{\sim}{y}\ |\ \underset{\sim}{Z}_i,\ \underset{\sim}{\psi}_g)f(\underset{\sim}{Z}\ |\ \underset{\sim}{\psi}_g))]. \\
&= \mathbb{E}^{\underset{\sim}{z}\ |\ \underset{\sim}{\psi}_g^{(k)}\ ,\ \underset{\sim}{y}}\left[\log\left(\prod_{i=1}^{n} f(y_i\ |\ \underset{\sim}{Z}_i\ ,\ \underset{\sim}{\theta})f(\underset{\sim}{Z}_i\ |\ \underset{\sim}{p}_{g-1})\right)\right] \\
&= \mathbb{E}^{\underset{\sim}{z}\ |\ \underset{\sim}{\psi}_g^{(k)}\ ,\ \underset{\sim}{y}}\left[\sum_{i=1}^{n}\left\{\left(\sum_{j=1}^{g-1} Z_{ij}\log f_j(y_i\ |\ \underset{\sim}{\theta}_j)\right)\right.\right. \\
&\quad \left.\left. + (1-\sum_{l=1}^{g-1} Z_{il})\log f_g(y_i\ |\ \underset{\sim}{\theta}_g)\right\}\right] \\
&\quad + \mathbb{E}^{\underset{\sim}{z}\ |\ \underset{\sim}{\psi}_g^{(k)}\ ,\ \underset{\sim}{y}}\left[\sum_{i=1}^{n}\left\{\left(\sum_{j=1}^{g-1} Z_{ij}\log p_j\right)\right.\right. \\
&\quad \left.\left. + (1-\sum_{j=1}^{g-1} Z_{ij})\log(1-\sum_{j=1}^{g-1} p_j)\right)\right\}\right] \\
&= \sum_{j=1}^{g-1}\left(\sum_{i=1}^{n} \mathbb{E}^{\underset{\sim}{z}\ |\ \underset{\sim}{\psi}_g^{(k)}\ ,\ \underset{\sim}{y}}[Z_{ij}]\log\ f_j(y_i\ |\ \underset{\sim}{\theta}_j)\right) \\
&\quad + \sum_{i=1}^{n}\left(1-\sum_{l=1}^{g-1}\mathbb{E}^{\underset{\sim}{z}\ |\ \underset{\sim}{\psi}_g^{(k)}\ ,\ \underset{\sim}{y}}[Z_{il}]\right)\log\ f_g(y_i\ |\ \underset{\sim}{\theta}_g) \\
&\quad + \sum_{j=1}^{g-1}\left(\log p_j \sum_{i=1}^{n}\mathbb{E}^{\underset{\sim}{z}\ |\ \underset{\sim}{\psi}_g^{(k)}\ ,\ \underset{\sim}{y}}[Z_{ij}]\right) \\
&\quad + \left(\log(1-\sum_{j=1}^{g-1} p_j)\right)\sum_{i=1}^{n}\left(1-\mathbb{E}^{\underset{\sim}{z}\ |\ \underset{\sim}{\psi}_g^{(k)}\ ,\ \underset{\sim}{y}}[Z_{ij}]\right),
\end{aligned}
$$

où

$$
\mathbb{E}^{\underset{\sim}{z}\ |\ \underset{\sim}{\psi}_g^{(k)}\ ,\ \underset{\sim}{y}}[Z_{ij}] = \mathbb{E}^{z_{ij}\ |\ \underset{\sim}{\psi}_g^{(k)}\ ,\ \underset{\sim}{y}}[Z_{ij}] = \frac{p_j^{(k)} f_j(x_i\ |\ \underset{\sim}{\theta}_j^{(k)})}{\sum_{r=1}^{g} p_r^{(k)} f_r(x_i\ |\ \underset{\sim}{\theta}_r^{(k)})}, \tag{2.2.1}
$$

et $\sum_{j=1}^{g} z_{ij} = 1 \quad \forall i = 1, \ldots, n.$ □

Nous devons poursuivre avec l'étape M, c'est-à-dire maximiser $Q(\underset{\sim}{\psi}\ ,\ \underset{\sim}{\psi}_g^{(k)})$ par rapport à $\underset{\sim}{\psi} = (p_1\ ,\ldots\ p_{g-1}\ \underset{\sim}{\theta}_1\ ,\ldots,\ \underset{\sim}{\theta}_g)$. Pour les paramètres $p_j, \quad j = 1, \ldots, g$, nous utilisons le théorème suivant.

Théorème 2.2.3. *Dans un mélange de lois à g composantes, pour j fixé, la maximisation de la probabilité p_j à la k^e itération est donnée par*

$$p_j^{(k+1)} = \frac{\sum_{i=1}^{n} \mathbb{E}^{z_{ij} \mid \underset{\sim}{\psi}_g^{(k)}, \underset{\sim}{y}}[Z_{ij}]}{n},$$

où $\mathbb{E}^{z_{ij} \mid \underset{\sim}{\psi}_g^{(k)}, \underset{\sim}{y}}[Z_{ij}]$ *est donnée à l'équation* (2.2.1).

DÉMONSTRATION. Posons

$$B_j = \sum_{i=1}^{n} \mathbb{E}^{z_{ij} \mid \underset{\sim}{\psi}_g^{(k)}, \underset{\sim}{y}}[Z_{ij}].$$

Donc, nous pouvons écrire

$$\begin{aligned} Q(\underset{\sim}{\psi}\ ,\ \underset{\sim}{\psi}_g^{(k)}) = &\sum_{j=1}^{g-1} \left(\sum_{i=1}^{n} \mathbb{E}^{z \mid \underset{\sim}{\psi}_g^{(k)}, \underset{\sim}{y}}[Z_{ij}] \log\ f_j(y_i \mid \underset{\sim}{\theta}_j) \right) \\ &+ \sum_{i=1}^{n} \left(1 - \sum_{j=1}^{g-1} \mathbb{E}^{z \mid \underset{\sim}{\psi}_g^{(k)}, \underset{\sim}{y}}[Z_{ij}] \right) \log\ f_g(y_i \mid \underset{\sim}{\theta}_g) \\ &+ \left(\sum_{j=1}^{g-1} (\log p_j) B_j + \log(1 - \sum_{j=1}^{g-1} p_j) B_g \right). \end{aligned}$$

Pour j fixé, nous avons

$$\begin{aligned} \frac{\partial}{\partial p_j} Q(\underset{\sim}{\psi}\ ,\ \underset{\sim}{\psi}_g^{(k)}) &= \frac{B_j}{p_j} - \frac{B_g}{1 - \sum_{l=1}^{g-1} p_l} \\ &= 0 \Leftrightarrow \frac{B_j}{p_j} = \frac{B_g}{p_g} \\ &\Leftrightarrow p_j = \frac{B_j p_g}{B_g}. \end{aligned}$$

Or, nous avons

$$\begin{aligned} 1 = \sum_{j=1}^{g-1} p_j + p_g &= \sum_{j=1}^{g-1} \frac{B_j p_g}{B_g} + p_g \\ &\Leftrightarrow p_g = \frac{B_g}{\sum_{j=1}^{g} B_j}, \end{aligned}$$

mais

$$\sum_{j=1}^{g} B_j = \sum_{i=1}^{n} \frac{\sum_{j=1}^{g} p_j f_j(y_i | \underset{\sim}{\theta}_j)}{\sum_{l=1}^{g} p_l f_l(y_i | \underset{\sim}{\theta}_l)} = n.$$

De façon générale, nous avons donc

$$\frac{\partial}{\partial p_j} Q(\underset{\sim}{\psi}\ ,\ \underset{\sim}{\psi}_g^{(k)}) = 0 \Leftrightarrow p_j = \frac{B_j}{n}.$$

□

Quant aux paramètres de $\underset{\sim}{\theta}_j$ de chacune des densités marginales, nous les trouvons en résolvant (voir proposition 2.1.2)

$$\frac{\partial Q(\underset{\sim}{\psi}\ ,\ \underset{\sim}{\psi}_g^{(k)})}{\partial \underset{\sim}{\theta}_j} = \sum_{i=1}^{n} \mathbb{E}^{z_{ij} \mid \underset{\sim}{\psi}_g^{(k)},\, \underset{\sim}{y}}[Z_{ij}] \frac{\partial}{\partial \underset{\sim}{\theta}_j} (\log f_j(y_i \mid \underset{\sim}{\theta}_j)) = 0,$$

où $\mathbb{E}^{z_{ig} \mid \underset{\sim}{\psi}_g^{(k)},\, \underset{\sim}{y}}[Z_{ig}] = 1 - \sum_{j=1}^{g-1} \mathbb{E}^{z_{ij} \mid \underset{\sim}{\psi}_g^{(k)},\, \underset{\sim}{y}}[Z_{ij}]$.

Notons que dans le cas plus particulier qui nous intéresse, $\log f_j(y_i \mid \underset{\sim}{\theta}_j)$ correspond au logarithme de la densité marginale pour une seule observation, c'est-à-dire

$$\log m_j(y_i \mid \underset{\sim}{\theta}_j) = \log[I_0(\kappa R_{ij}^*)] - \log[2\pi] - \log[I_0(\kappa)] - \log[I_0(\omega_j \kappa)].$$

2.2.2.4. *Application de l'algorithme aux données de l'exemple*

Nous avons remarqué que la méthode ML-II décrite à la section 1.1.2 ajuste une courbe unimodale aux données. En considérant non pas une seule loi marginale mais plutôt un mélange de lois marginales ayant comme mode chacun des sommets présents sur la courbe des données, nous espérons améliorer l'ajustement. Nous avons ici appliquer l'algorithme EM à un mélange de $g = 2$ lois marginales car nous retrouvons uniquement 2 modes sur l'histogramme de la figure 1.2.2. Nous fournissons au tableau 2.2.1 les valeurs initiales considérées de même que les estimateurs pour chacun des paramètres du mélange. De plus, nous avons ajouter l'heure correspondante pour chacun des μ_j impliqués dans le modèle considéré. Pour ce faire, nous avons utilisé la formule

$$\left[\frac{12y}{\pi}\right] + 60\left(\frac{12y}{\pi} - \left[\frac{12y}{\pi}\right]\right)$$

qui nous permet de transformer des valeurs y (en radians) en heure de la journée. Il est à noter que dans cette formule, $[x]$ est la partie entière de x.

Pour les valeurs initiales, nous avons choisi $g-1$ limites sur abscisse, notée $L_1, \ldots, L_{g-1}$, à partir de l'histogramme, qui nous semblait départager les g densités du mélange. Nous avons déterminé la probabilité initiale $p_{j-1}^{(0)}, \quad j = 2, \ldots, g$, simplement par $p_{j-1}^{(0)} = \text{card}(L_{j-2} < v \leq L_{j-1})/n$, avec $L_0 = 0$ et $p_g = 1 - \sum_{j=1}^{g-1} p_g$ et n le nombre de données considérés. Pour obtenir les estimateurs initiaux $\mu_j^{(0)}$, nous choisissons les modes de l'histogramme, et finalement pour $\omega_j^{(0)}$, nous utilisons l'approximation donnée à la proposition 1.2.1 pour chacun des sous-échantillons définis par les limites $L_j, \quad j = 1, \ldots, g-1$. Enfin, la valeur $\kappa^{(0)}$ est choisie de façon arbitraire mais près de un afin de minimiser son influence à l'étape initiale.

Pour illustrer l'ajustement, nous avons ajouté la figure 2.2.1 qui permet de comparer l'histogramme des données, l'estimation par une loi marginale pour un seul mode et par un mélange de deux lois marginales avec les valeurs du tableau 2.2.1.

2.3. Sélection du modèle d'ajustement

Il arrive fréquemment que nous soyons confronté au problème de déterminer un modèle d'ajustement à des données. Plusieurs critères sont proposés dans la littérature, mais nous avons choisi de considérer les trois critères suivants : maximisation de l'entropie, critères AIC et BIC.

Définition 2.3.1. *Nous définissons l'entropie associé à la fonction de densité $f(y \mid \underline{\theta})$, notée $H_f(\underline{\theta})$ la fonction*

$$H_f(\underline{\theta}) = -\mathbb{E}^{f(y \mid \underline{\theta})}[\log f(y \mid \underline{\theta})].$$

Par la loi faible des grands nombres, nous pouvons estimer $H_f(\underline{\theta})$ par

$$\widehat{H}_f(\underline{\theta}) = -\frac{1}{n}\sum_{i=1}^{n} \log f(y_i \mid \hat{\underline{\theta}}),$$

TABLEAU 2.2.1. Estimation des paramètres d'une seule densité marginale et d'un mélange à deux composantes pour les données de l'exemple

	une composante		deux composantes	
	Valeur initiale	**Valeur finale**	**Valeur initiale**	**Valeur finale**
p_1			0,3	0,288
p_2			0,7	0,712
μ_1	3,648 (13h56)	3,611 (13h48)	1,0 (3h49)	1,031 (3h56)
μ_2			4,0 (15h17)	3,943 (15h06)
ω_1	1,200	0,632	0,8	457,857
ω_2			0,8	1,368
κ	0,278	1,000	1,0	1,522

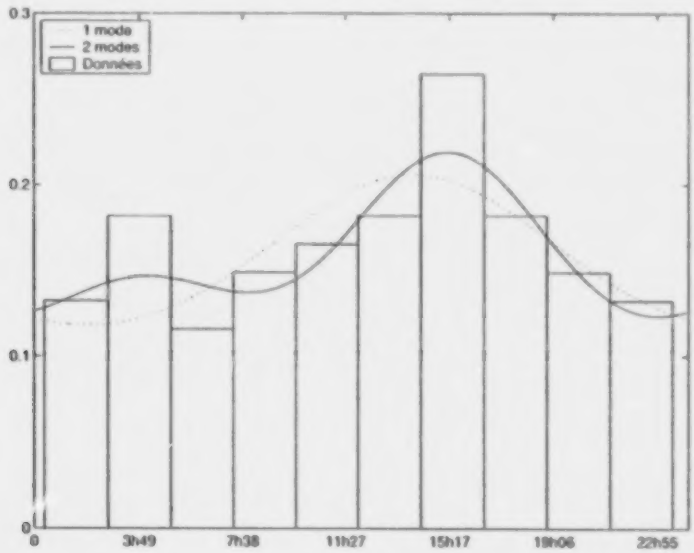

FIGURE 2.2.1. Comparaison de la répartition des données de l'exemple avec une approximation par une densité marginale et un mélange de deux densités marginales

où $\hat{\underline{\theta}}$ est l'estimateur du maximum de vraisemblance de $\underline{\theta}$ obtenu à partir de $y_1, \ldots, y_n$.

L'entropie est en fait une quantité qui nous renseigne sur l'information que la densité $f(y \mid \underline{\theta})$ nous apporte sur y. Une première option pour le choix du modèle serait de choisir celui qui maximise $-\widehat{H}_f(\underline{\theta})$. Toutefois, comme ce critère ne tient pas compte du nombre de paramètres estimés pour chacun des modèles, Aikaike (1974) a suggéré un critère qui tient compte du nombre de paramètres de chacun des modèles.

Définition 2.3.2. *Le critère AIC consiste à maximiser la fonction de vraisemblance $L_j(\underline{\theta}_j \mid \underline{y})$ pour chacun des j modèles candidats, et ensuite choisir le modèle pour lequel $\log L_j(\underline{\theta}_j \mid \underline{y}) - k_j$ est le plus élevé, où k_j est la dimension du vecteur de paramètres θ_j.*

Schwarz (1978) propose une alternative de ce critère AIC : le critère d'information de Bayes (BIC).

Définition 2.3.3. *Soit $Y_i, \quad i = 1, \ldots, n$ n variables aléatoires de fonction de vraisemblance $L(\underline{y} \mid \underline{\theta})$ où $\underline{\theta}$ est le vecteur des k paramètres. Supposons que nous devons choisir parmi j modèles. Le critère d'information de Bayes (BIC) stipule de choisir le modèle j tel que*

$$\log L_j(\underline{y} \mid \underline{\theta}_j) - \frac{k_j}{2} \log\left(\frac{n}{2\pi}\right)$$

est maximum.

Le critère BIC ne prend pas uniquement en considération le nombre de paramètres inclus dans le modèle mais aussi le nombre de données dans l'échantillon. En effet, le critère AIC pénalise en quelque sorte pour le nombre de paramètres dans le modèle alors que le BIC pondère cette pénalité par un facteur dépendant

de la taille de l'échantillon.

2.3.1. Application des critères de sélection de modèle

Notre but ici est de statuer sur le modèle à conserver pour ajuster les données de l'exemple. Nous avons donc fourni au tableau 2.3.1 les résultats des différents critères pour chacun des modèles considérés, c'est-à-dire les modèles unimodal et bimodal.

TABLEAU 2.3.1. Comparaison des différents modèles pour les données de l'exemple

Modèle	$-\widehat{H}_f(\underset{\sim}{\theta})$	**AIC**	**BIC**
1 mode	-1,819	-4,819	-5,970
2 modes	-1,809	-7,809	-10,111

Par l'observation de la figure 2.2.1 qui compare les courbes des deux modèles à l'histogramme des données, nous constatons que le modèle à deux modes semble mieux s'ajuster à l'histogramme. Nous voyons que le critère basé sur l'entropie désigne le modèle à deux modes comme celui à considérer. Toutefois, les deux autres critères optent pour le modèle unimodal car ils pénalisent sévèrement l'action d'ajouter un mode supplémentaire au modèle. Toutefois, comme nous savons que les données de cet exemple ont été obtenues à partir d'un mélange de lois de von Mises

$$f(y_i \mid \underset{\sim}{\theta}) = 0{,}35 \times f_1(y_i \mid \underset{\sim}{\theta}_1) + 0{,}65 \times f_2(y_i \mid \underset{\sim}{\theta}_2),$$

où $y_i \mid \underset{\sim}{\theta}_1 \sim \text{VM}(1\ ,\ 0{,}8)$ et $y_i \mid \underset{\sim}{\theta}_2 \sim \text{VM}(4\ ,\ 0{,}8)$, le modèle à deux composantes était celui recherché. Rappelons que le modèle bimodal obtenu pour l'ajustement à l'aide de l'algorithme EM est donné par

$$f_{\text{ajustement}}(y_i \mid \underset{\sim}{\theta}) = 0{,}288 \times m_1(y_i \mid \underset{\sim}{\theta}_1) + 0{,}712 \times m_2(y_i \mid \underset{\sim}{\theta}_2),$$

où $y_i \mid \underset{\sim}{\theta}_j \quad j = 1, 2$ sont des lois marginales telles que définies à l'équation (2.1.3) avec comme paramètres respectifs $\mu_1 = 1{,}03$, $\omega_1 = 457{,}857$, $\mu_2 = 3{,}94$, $\omega_2 = 1{,}368$ et $\kappa = 1{,}522$.

Cet exemple nous montre que le critère de sélection de modèle qui est à privilégier dans notre situation est le maximum de l'entropie. Les autres critères sont à utiliser avec parcimonie compte tenu du fait que l'ajout d'une composante au modèle augmente chaque fois de trois le nombre de paramètres à estimer.

Nous avons jusqu'ici modéliser la répartition de données dans le temps. Dans le prochain chapitre, nous allons ajuster une courbe sur des données dans le temps afin d'obtenir une densité qui nous sera utile dans le calcul d'un rapport de cote ou d'un risque relatif.

Chapitre 3

ÉTUDE ÉPIDÉMIOLOGIQUE ET RÉSULTATS

3.1. Présentation de l'étude

Avec la forte croissance de l'utilisation des téléphones mobiles/cellulaires, nous pouvons être tentés de vérifier si l'usage de ces appareils pendant la conduite automobile engendre un risque plus élevé d'accidents de la route. Des études sur simulateurs portant sur peu de sujets (entre 12 et 27 sujets) ont montré plusieurs effets potentiellement néfastes pour la conduite automobile (Stein, Parseghian et Allen, 1987; McKnight et McKnight, 1993; Alm et Nilsson, 1995). L'analyse de données d'une autre étude épidémiologique, portant cette fois sur 699 utilisateurs impliqués dans un accident avec dommages matériels sérieux, arrive à la conclusion que le risque d'accident est environ 4 fois plus élevé si un conducteur fait usage de son téléphone tout en conduisant (voir Redelmeier et Tibshirani, 1997).

Plus récemment, une équipe de chercheurs du C.R.T. a fait une étude épidémiologique comportant deux cohortes suffisamment nombreuses, les utilisateurs du téléphone mobile et les non-utilisateurs, pour vérifier s'il existe une réelle association entre l'utilisation du téléphone pendant la conduite automobile et les accidents de la route tout en considérant certains regroupements de conducteurs et certains types d'accidents.

Au niveau méthodologique, la Société de l'Assurance Automobile du Québec (SAAQ) a fait, en novembre 1999, l'envoi massif d'un questionnaire et d'une lettre

de consentement à 175 000 titulaires de permis de conduire de classe 5 (véhicule de promenade), stratifiés selon l'âge et le sexe tout en respectant les spécifications de l'équipe de recherche du C.R.T. Le questionnaire portait surtout sur les habitudes de conduite (fréquence d'utilisation du véhicule après 20 heures, écoute de musique durant la conduite, ...) et l'exposition au risque d'accident de la route (nombre de kilomètres parcourus au cours des 12 derniers mois, etc). Quant à la lettre de consentement, elle demandait l'autorisation des participants pour consulter leur dossier de conduite de la SAAQ (date et heure des accidents, points d'inaptitude, etc.), et pour obtenir les données de leur compagnie de téléphonie (Bell mobilité, Cantel, Fido ou Clearnet) sur l'utilisation de leur téléphone cellulaire (début et fin de chaque appel fait ou reçu, appels d'urgence, etc.).

Les méthodes statistiques qui ont été utilisées par l'équipe de recherche sont des tests du khi-deux, des calculs de risques relatifs à partir de tableaux de contingence et finalement des modèles de régression logistique-normale pour estimer la force des liens entre les variables explicatives et les accidents. Globalement, leurs résultats ont montré qu'un utilisateur de téléphone cellulaire a 38% plus de risque d'avoir un accident qu'un non-utilisateur. Plus spécifiquement, pour les seuls utilisateurs, le rapport de cotes augmente environ à 2 pour ceux faisant au moins 135 appels par mois.

Ils ont démontré que le facteur de l'âge n'a pas d'influence sur le taux d'accidents uniquement pour le groupe des utilisatrices. Pour tous les autres regroupements, ce facteur est à considérer. Chez les seules utilisatrices, le nombre d'appels reçus est déterminant quant au risque d'accident de voiture. En effet, celles qui reçoivent 55 appels et plus par mois sont 2,3 fois plus à risque d'avoir un accident que celles qui reçoivent aucun ou un appel par mois.

Quant aux seuls utilisateurs, l'âge a un effet sur le risque. À preuve, l'équipe de recherche a démontré que le groupe des conducteurs âgés entre 16 et 24 ans ont 2,6 fois plus de risque que leurs pairs de 55 à 64 ans (Laberge-Nadeau et al.,

2001, p.112). Un autre facteur important est celui de la fréquence d'utilisation du téléphone mobile. Ceux qui font plus de 135 appels par mois ont deux fois plus de risque d'avoir un accident que ceux faisant moins de 10 appels. Il est intéressant de remarquer que les faibles utilisateurs ont le même risque d'avoir un accident que les non-utilisateurs. Mentionnons que les risques calculés tiennent compte de l'âge et du nombre de kilomètres parcourus par le conducteur.

Si nous nous intéressons spécifiquement aux accidents qui engendrent des blessures corporels, seules les variables du nombre de kilomètres parcourus annuellement et du nombre d'appels reçus par mois servent à expliquer le risque d'accidents. Ces résultats statistiques ont été obtenus suite à l'analyse de données fournies par une vaste enquête épidémiologique auprès d'un échantillon décrit à la section suivante.

3.1.1. Information sur le jeu de données

La population de départ est constituée de l'ensemble des détenteurs de permis de conduire de classe 5 (véhicule de promenade), habitant une des 34 villes du Québec ayant un revenu moyen par ménage supérieur à 30 000$. Cette contrainte sur le revenu dans la sélection de l'échantillon a simplement servi à maximiser le nombre d'utilisateurs de téléphone cellulaire. Malgré leur revenu moyen par ménage inférieur à 30 000$, les villes de Montréal, Québec et Laval ont été incluses étant donné leur important poids démographique par rapport à l'ensemble de la province.

Autre critère important : la position géographique de la ville. En effet, comme le service de téléphonie cellulaire n'était pas encore fonctionnel dans certaines villes situées au nord du Québec, uniquement les villes desservies par le service ont été retenues. Enfin, lors d'une étude pilote, les chercheurs du C.R.T. ont remarqué qu'environ 7% des hommes répondaient au questionnaire alors que ce chiffre passait à 38% pour les femmes. Les hommes ont donc été sur-échantillonnés

(deux hommes pour une femme et ce dans chacune des villes) afin d'avoir suffisamment d'hommes répondants. En tenant compte du nombre de répondants selon le sexe, nous obtenons un taux global de réponse de 21%. Toujours par cette étude pilote, nous avons obtenu une estimation du taux d'utilisation du téléphone mobile de 23%.

Il a été démontré par les chercheurs du C.R.T. que pour atteindre une puissance statistique suffisante afin de déceler un risque d'accident, un échantillon de 175 000 titulaires de permis de classe 5 était nécessaire. En effet, les chercheurs ont montré que si nous voulons détecter un risque relatif de 1,75 avec une puissance $1-\beta$ égale à 95% et un niveau α de 1% pour le test bilatéral pour l'hypothèse H_0, l'usage du téléphone cellulaire en conduisant n'a aucun effet, contre l'alternative H_1, l'usage du téléphone cellulaire en conduisant a un effet quelconque, alors nous aurons besoin de $N = 8587$ utilisateurs du téléphone mobile. Cette estimation est basée sur la formule suivante proposée par Schlesselman (1974) :

$$N = \frac{\left(Z_{\alpha/2} \sqrt{2u(1-u)} + Z_{\beta} \sqrt{f(1-f) + p_3(1-p_3)} \right)^2}{(f-p_3)^2}, \qquad (3.1.1)$$

où $Z_{\alpha/2}$ est tel que $\mathbb{P}(Z > Z_{\alpha/2}) = \alpha/2$, Z_{β} est tel que $\mathbb{P}(Z > Z_{\beta}) = \beta$, $u = \frac{1}{2} f \left(1 + \frac{RR}{1+f(RR-1)}\right)$, RR est le risque relatif à déceler, f est la probabilité d'avoir un accident et $p_3 = \frac{f\,RR}{1+f(RR-1)}$.

Nous avons fourni à la figure 3.1.1, un graphique sur lequel nous pouvons voir la taille échantillonnale nécessaire en fonction de f, la probabilité d'avoir un accident pour des valeurs entre 0,01 et 0,1. Cette courbe se base sur les hypothèses suivantes : $RR = 1{,}75$, $\alpha = 0{,}01$ et $\beta = 0{,}05$.

Donc si un envoi est fait à 175 000 personnes, nous pouvons estimer qu'il y aura 36 750 répondants (175 000 × 21%). Comme nous avons fait l'estimation que 23% de ces 36 750 répondants sont des utilisateurs du téléphone mobile, il y aurait donc 8 453 personnes qui est environ le nombre d'utilisateurs nécessaires pour déceler un risque d'accident de 1,75 avec une puissance de 95% et un niveau

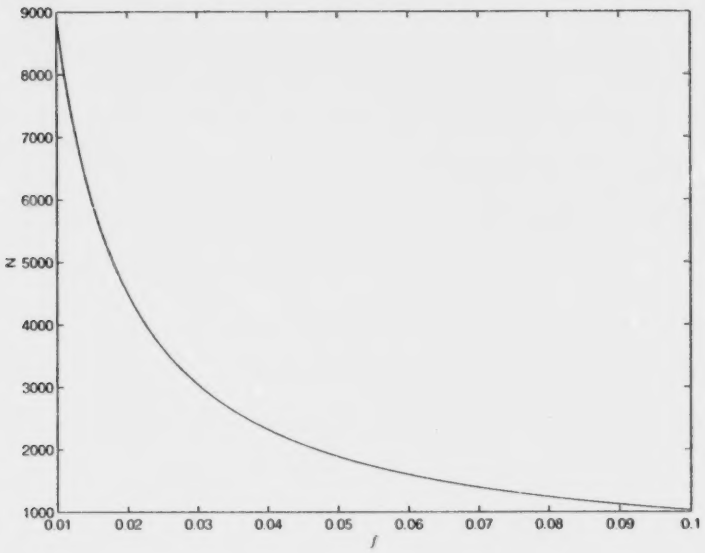

FIGURE 3.1.1. Courbe de la taille de l'échantillon (N) en fonction de f selon l'équation (3.1.1)

de 1%. L'échantillon a été sélectionné de façon à maximiser le nombre d'utilisateurs et d'utilisatrices de téléphone mobile. Une étude de clientèle a été faite à partir d'une source d'information (CWTA, 1998). Nous y apprenons qu'environ 9% des utilisateurs de téléphone mobile sont âgés entre 15 et 24 ans et 78% entre 25 et 54 ans. Nous avons donc, en plus de sur-échantillonner les hommes, respecté le pourcentage d'utilisation du téléphone mobile pour chacune des classes d'âge afin de maximiser le nombre de répondants faisant usage d'un téléphone cellulaire. Nous pouvons voir au tableau 3.1.1 la répartition des 175 000 titulaires de permis selon le sexe et l'âge.

Le questionnaire a été complété par 38 300 personnes dont 36 079 ont aussi signé la lettre de consentement. Ce sont ces derniers qui formeront l'échantillon de notre étude. Nous pouvons voir au tableau 3.1.2 le taux de réponse selon l'âge et le sexe. Nous remarquons que le taux global de réponse pour les femmes est, tel que prévu par l'étude pilote, un peu plus élevé. Autre fait intéressant à relever est celui que le taux de réponse des hommes croît avec l'âge tandis que chez les femmes, il est passablement constant.

TABLEAU 3.1.1. Répartition des 175 000 titulaires de permis selon l'âge et le sexe

	Hommes		Femmes		Ensemble	
Groupe d'âge	**N**	**%**	**N**	**%**	**N**	**%**
16-24 ans	10 629	9,1	5 248	9,0	15 877	9,1
25-34 ans	28 409	24,4	14 576	25,0	42 985	24,6
35-44 ans	38 671	33,1	19 244	33,0	57 915	33,1
45-54 ans	28 344	24,2	14 001	24,0	42 345	24,2
55-64 ans	10 630	9,1	5 248	9,0	15 878	9,1
Total	**116 683**	**100,0**	**58 317**	**100,0**	**175 000**	**100,0**

TABLEAU 3.1.2. Taux de réponse selon l'âge et le sexe

	Sexe des répondants					
	Hommes			Femmes		
Groupes	**Répondants**	**N**	**taux**	**Répondants**	**N**	**taux**
16-24 ans	1 412	10 629	13,3	1 158	5 248	22,1
25-34 ans	4 169	28 409	14,7	2 886	14 576	19,8
35-44 ans	7 322	38 671	18,9	4 253	19 244	22,1
45-54 ans	6 973	28 344	24,6	3 478	14 001	24,8
55-64 ans	3 068	10 630	28,9	1 360	5 248	25,9
Total	**22 944**	**116 683**	**19,7**	**13 135**	**58 317**	**22,5**

Il est important de noter que tous les individus sélectionnés pour l'échantillon n'avaient pas nécessairement leur permis de conduire au cours des années 1996 à 1999. Nous retrouvons dans le tableau 3.1.3, le nombre de titulaires ayant leur permis de conduire de classe 5 au 31 décembre de l'année en cours. Toutefois, 2% des 166 378 titulaires de permis de conduire classe 5 ont obtenu leur permis au cours de l'année 1996 dont 50% d'entre eux l'ont obtenu après le 30 juin, ayant ainsi 6 mois ou moins d'expérience de conduite en 1996. Pour obtenir le nombre moyen d'accidents par année, nous avons donc fait un calcul pondéré, basé sur le nombre de mois d'expérience de conduite pour l'année considérée, variant de 1 à

TABLEAU 3.1.3. Nombre de titulaires détenant un permis de conduire de classe 5 au 31 décembre de l'année en cours

Année	Hommes	Femmes	Ensemble
1996	110 995	55 383	166 378
1997	113 511	56 736	170 247
1998	115 080	57 498	172 578
1999	116 683	58 317	175 000

12 mois pour chaque individus. Ainsi, pour l'année 1996, au lieu d'avoir 110 995 titulaires masculins, nous obtenons 109 801,3 titulaires/année masculins. En tenant compte de cette information, nous pouvons construire le tableau 3.1.4 qui nous donne le nombre moyen d'accidents annuels pour 1000 titulaires de l'échantillon selon le sexe et l'année. Globalement, nous remarquons qu'il y a une diminution du nombre moyen d'accidents à travers les années, et ce tant pour les hommes que pour les femmes. Plus précisément, cette moyenne d'accidents est 1,56 fois moins élevée chez les femmes (26,8 pour les femmes contre 41,8 pour les hommes).

À des fins de comparaison, nous avons ajouté le tableau 3.1.5 qui montre les mêmes mesures mais pour 1 000 titulaires de la population du Québec âgés entre 16 et 64 ans. Nous voyons que les résultats générés par notre échantillon sont toujours plus faibles que dans la population. Il sera important de tenir compte de ce fait lors de l'interprétation de nos résultats. À toute fin pratique, les résultats obtenus vont sous-estimer les valeurs de la population.

Dans le tableau 3.1.6, nous dressons un portrait du nombre d'accidents de la route dans lesquels les participants ayant remplis le questionnaire ont été impliqués au cours des 24 mois précédents novembre 1999. Nous pouvons remarquer dans ce tableau que la répartition des accidents est similaire selon le sexe alors que le tableau 3.1.4 indique que les hommes ont plus d'accidents en moyenne par année. Nous pouvons donc penser que les hommes de l'échantillon de 175 000 personnes ayant eu beaucoup d'accidents de voiture n'ont pas toujours complété

TABLEAU 3.1.4. Nombre moyen d'accidents annuels pour 1000 titulaires de l'échantillon selon le sexe et l'année entre 1996 et 1999.

	Hommes		Femmes	
Période d'observation	**N**	**Accidents**	**N**	**Accidents**
1996	109 801,3	44,1	54 708,4	29,1
1997	112 302,7	44,2	56 059,1	27,5
1998	114 293,4	40,7	57 100,4	26,9
1999	116 025,8	38,7	57 967,7	24,3
Moyenne		41,8		26,8

TABLEAU 3.1.5. Nombre moyen d'accidents annuels pour 1000 titulaires de la population du Québec âgés entre 16 et 64 ans selon le sexe et l'année.

	Hommes		Femmes	
Période d'observation	**N**	**Accidents**	**N**	**Accidents**
1996	2 061 265	69,9	1 836 959	36,9
1997	2 061 902	72,6	1 852 555	38,1
1998	2 071 474	65,7	1 860 056	35,3
1999	2 083 934	61,1	1 871 032	32,7
Moyenne		67,3		35,7

le questionnaire et donc ne sont pas considérés dans le tableau 3.1.6.

Nous sommes intéressé à ces données sur les accidents car nous voulons ultimement vérifier s'il existe un risque plus élevé d'avoir un accident de voiture tout en utilisant son téléphone cellulaire. Pour ce faire, nous devons avoir recours à la théorie sur les données catégorielles présentée au premier chapitre pour construire un tableau de contingence qui nous permettra de calculer un rapport de cotes moyen ainsi qu'un risque relatif moyen.

TABLEAU 3.1.6. Nombre d'accidents dans lequel un participant a été impliqué de novembre 1997 à novembre 1999 selon le sexe et son usage du téléphone cellulaire

	Hommes		Femmes	
Nombre d'accidents	**Utilisateurs** (%)	**Non-util.** (%)	**Utilisateurs** (%)	**Non-util.** (%)
0 accident	83,8	86,7	83,8	88,0
1 accident	14,1	12,1	14,3	11,0
2 accidents ou plus	2,13	1,21	1,96	1,03
Nombre	9 313	13 515	3 321	9 727

3.2. MODÉLISATION DES PROBABILITÉS MARGINALES

Dans ce mémoire, nous voulons ultimement tenter de quantifier le degré d'association entre deux variables aléatoires catégorielles qui sont X, "faire l'usage du téléphone mobile en conduisant" pour les seuls utilisateurs et Y, "avoir un accident de voiture (tous types confondus)". Nous avons vu au premier chapitre que le calcul du rapport de cotes ou du risque relatif pour mesurer le niveau d'association s'effectue à partir des différents effectifs échantillonnaux n_{ij} ou des probabilités observées $\hat{\nu}_{ij}$, $\quad i, j = 1, 2$.

Toutefois, nous ne disposons pas de toutes ces données. Pourtant, nous allons réussir à estimer ces deux mesures à l'aide du tableau de contingence 3.2.1 où n_{11}, n_{21} et n sont des effectifs connus (par minute) tandis que $\Upsilon 1$ et $\Upsilon 2$ sont des densités de probabilités marginales desquelles nous échantillonnerons des probabilités. Cependant afin d'obtenir ces différents paramètres, nous devons examiner les données qui ont été recueillies.

3.2.1. Données disponibles

L'enquête réalisée par le C.R.T. a permis de bâtir deux jeux de données qui peuvent être mis en relation par un numéro d'identification pour chaque individu. Dans un premier temps, la SAAQ nous a transmis une liste de tous les accidents

TABLEAU 3.2.1. Tableau de contingence 2 × 2 modifié.

X / Y	Oui	Non	Total
Oui	n_{11}		$\Upsilon 1$
Non	n_{21}		
Total	$\Upsilon 2$		n

rapportés par un policier où un individu participant à l'étude a été impliqué pour la période du premier août 1998 au 31 août 2000. Dans un deuxième temps, les quatres compagnies de téléphonie ont fourni un fichier contenant tous les appels faits ou reçus par un individu possédant un téléphone cellulaire et participant à l'étude dans la même période.

Nous allons d'abord nous intéresser au jeu de données sur les appels. Pour chaque utilisateur et pour chacun des appels, nous possédons l'heure exacte (à la seconde près) où la conversation téléphonique a débuté et s'est terminée. Nous sommes donc en mesure d'obtenir, pour l'ensemble des utilisateurs, le nombre de téléphones mobiles en fonction pour chacune des minutes de la journée. Toutefois, le nombre obtenu pour chaque minute correspond en fait à un nombre d'appels/minute, car si un seul appel dure 6 minutes alors il sera considéré comme 6 appels d'une minute chacun ou encore 6 appels/minute.

D'autres informations pertinentes sont également disponibles dans le jeu de données sur les accidents. Pour chaque accident, nous possédons l'heure à laquelle il s'est produit telle qu'inscrite sur le rapport de police complété par l'agent. Toutefois, certaines études ont démontré que cette mesure était imprécise (voir Baker, 1971) et pouvait varier de 2 à 30 minutes du moment exact. Cette heure inscrite sur le rapport doit donc être vu comme une borne supérieure du moment où l'accident s'est véritablement produit. En effet lorsqu'un accident survient, un usager va faire plusieurs appels soit pour alerter la police, obtenir un remorquage, aviser ses proches, etc. Nous voulons éviter de considérer ces appels faits après le moment de l'accident comme des appels ayant potentiellement causés l'accident.

Nous avons donc choisi de considérer que l'accident est survenu à un moment quelconque dans un intervalle de 15 minutes précédant l'heure inscrite sur le rapport. L'usage d'une fenêtre de temps pour déterminer le moment de l'accident fait en sorte que nous nous retrouvons dans le même cadre que pour les appels, c'est-à-dire que chaque accident comptera pour 15 accidents/minute. Grâce à ces nombres d'accidents ou d'appels pour chaque minute, nous pouvons modéliser leur répartition dans une journée.

3.2.2. Modélisation de la répartition des appels et des accidents

Comme nous nous intéressons maintenant à des données par minute dans l'intervalle d'une journée, nous sommes dans un contexte circulaire tel que présenté au premier chapitre. Nous allons donc modéliser la variable aléatoire X, "faire l'usage du téléphone mobile", sous les mêmes hypothèses que la proposition 2.1.1. En somme, nous supposons que $X_i, \quad i = 1, ..., n_{\text{app}}$ sont n_{app} variables aléatoires indépendantes de densité $\text{VM}(\mu_i, \kappa)$ avec $\mu_i \sim \text{VM}(\mu_0, \omega\kappa)$. Ainsi l'estimation de la densité marginale par la méthode ML-II est donnée par

$$m(\underset{\sim}{x}) = \frac{\prod_{i=1}^{n_{\text{app}}} I_0(\hat{\kappa} R_i^*)}{(2\pi)^{n_{\text{app}}} I_0^{n_{\text{app}}}(\hat{\kappa}) I_0^{n_{\text{app}}}(\hat{\omega}\hat{\kappa})},$$

où $R_i^* = \sqrt{C_i^{*2} + S_i^{*2}}$, $C_i^* = \cos x_i + \hat{\omega} \cos \hat{\mu}_0$, $S_i^* = \sin x_i + \hat{\omega} \sin \hat{\mu}_0$ et $\hat{\mu}_0$, $\hat{\omega}$, $\hat{\kappa}$ tels que définis à la proposition 2.1.2.

Nous avons fourni le tableau 3.2.2 qui donne l'estimation des paramètres impliqués dans le modèle, et la figure 3.2.1 qui montre la répartition des appels ainsi que son estimation par la méthode ML-II. À noter que les valeurs initiales ont été choisies par la méthode présentée à la section 2.2.2.4. Toutefois, nous voyons clairement que la courbe présente une allure multimodale. Nous allons donc appliquer l'algorithme Espérance-Maximisation (voir section 2.2.2) à un mélange de lois marginales définies à la proposition 2.1.1 afin de tenter d'ajuster une meilleure courbe qui tiendrait compte des différents modes. Les résultats de l'estimation des divers paramètres pour les différents modèles sont inclus dans le tableau 3.2.3,

TABLEAU 3.2.2. Estimation des paramètres par la méthode ML-II pour les appels

	Valeur initiale	**Valeur finale**
μ_1	3,763 (14h22)	3,764 (14h23)
ω_1	0,99	0,132
κ	0,059	0,999

et la figure 3.2.2 permet de comparer l'estimation des courbes pour chacun des modèles considérés. Il est à noter que les courbes pour un et deux modes sont confondues.

Nous devons maintenant statuer sur le modèle à conserver en utilisant les critères décrits à la section 2.3. Les résultats des différents modèles sont présentés au tableau 3.2.4. Le critère du maximum d'entropie ne nous permet pas de conclure. En considérant les deux autres critères, nous concluons que le modèle à conserver est celui possédant un seul mode puisqu'il maximise les valeurs obtenues pour les critères AIC et BIC. C'est donc dire que la densité que nous utiliserons pour ajuster les données des appels sera donnée par :

$$f_{\text{app}}(X \mid \underline{\theta}) = m(X \mid \underline{\theta}),$$

où $m(X \mid \underline{\theta})$ est défini à la proposition 2.1.1 et $\underline{\theta} = (\mu,\ \omega,\ \kappa)$ est donné au tableau 3.2.2.

Nous allons poursuivre en modélisant la variables Y, "avoir un accident de voiture" (tous types confondus), toujours sous les mêmes hypothèses que la proposition 2.1.1. Les estimations des paramètres trouvées par la méthode ML-II sont données au tableau 3.2.5, et la figure 3.2.3 nous permet de comparer la courbe des données à celle obtenue par la méthode ML-II. Encore une fois, nous pouvons remarquer que la courbe des données n'est pas unimodale. Tout comme pour les données sur les appels, nous allons tenter d'améliorer l'ajustement en appliquant

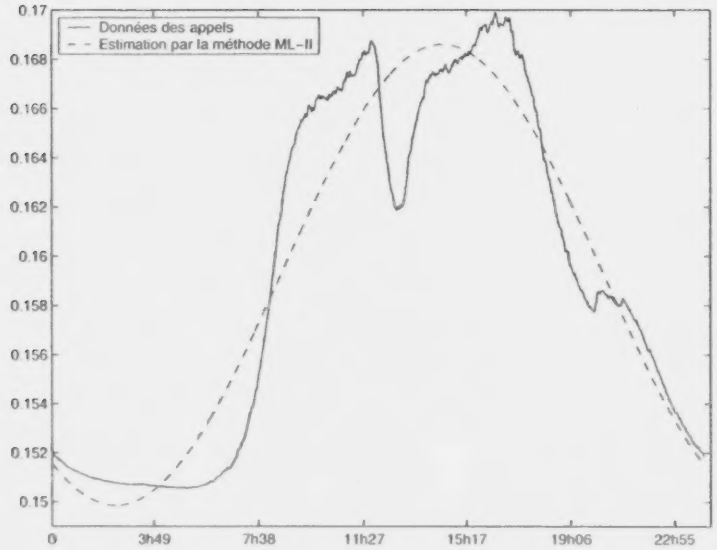

FIGURE 3.2.1. Comparaison de la répartition des appels selon l'heure de la journée et son approximation par la méthode ML-II.

l'algorithme Espérance-Maximisation à un mélange de lois marginales. Les résultats des estimations des paramètres impliqués dans chacun des modèles sont fournis au tableau 3.2.6 pour les modèles avec deux et trois composantes et au

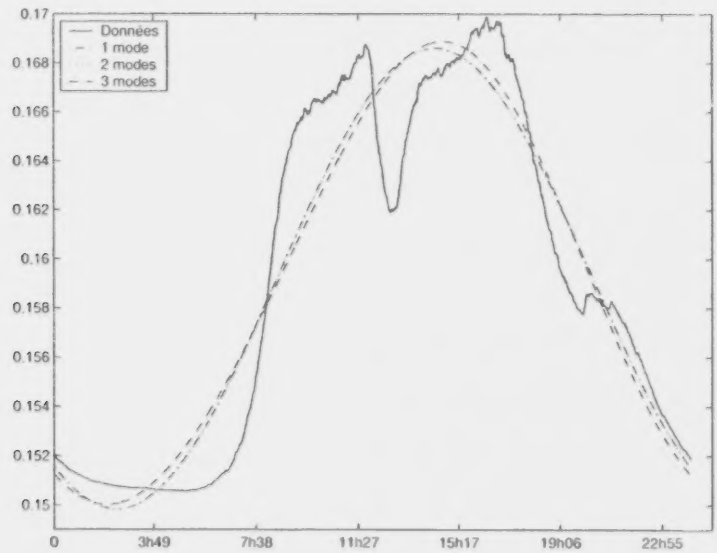

FIGURE 3.2.2. Comparaison de la répartition des appels selon l'heure de la journée et son approximation selon le nombre de modes.

TABLEAU 3.2.3. Estimation des paramètres d'un mélange à deux ou trois composantes pour les données sur les appels

	Deux composantes		Trois composantes	
	Valeur initiale	**Valeur finale**	**Valeur initiale**	**Valeur finale**
p_1	0,508	0,508	0,508	0,507
p_2	0,492	0,492	0,291	0,292
p_3			0,201	0,201
μ_1	1,672 (6h23)	3,185 (12h10)	1,672 (6h23)	1,884 (7h12)
μ_2	4,735 (18h05)	4,116 (15h43)	4,123 (15h45)	4,092 (15h38)
μ_3			5,636 (21h32)	5,434 (20h45)
ω_1	1,553	1,242	1,553	1,359
ω_2	1,792	2,090	4,433	5,098
ω_3			1,707	1,600
κ	0,2	0,287	0,400	0,405

TABLEAU 3.2.4. Comparaison des différents modèles pour les données sur les appels

Modèle	$-\widehat{H}_f(\underline{\theta})$	**AIC**	**BIC**
1 mode	-1,837	-4,837	-29,069
2 modes	-1,837	-7,837	-56,302
3 modes	-1,837	-10,837	-83,534

tableau 3.2.7 pour ceux à quatre et cinq composantes. La figure 3.2.4 permet de comparer l'ajustement des différents modèles aux données sur les accidents.

Pour décider du meilleur modèle pour l'ajustement des données des accidents, nous allons encore une fois utiliser les critères proposés à la section 2.3 (voir

TABLEAU 3.2.5. Estimation des paramètres par la méthode ML-II pour les accidents

	Valeur initiale	**Valeur finale**
μ_1	3,886 (14h51)	3,878 (14h49)
ω_1	1,070	2,938
κ	0,837	1,096

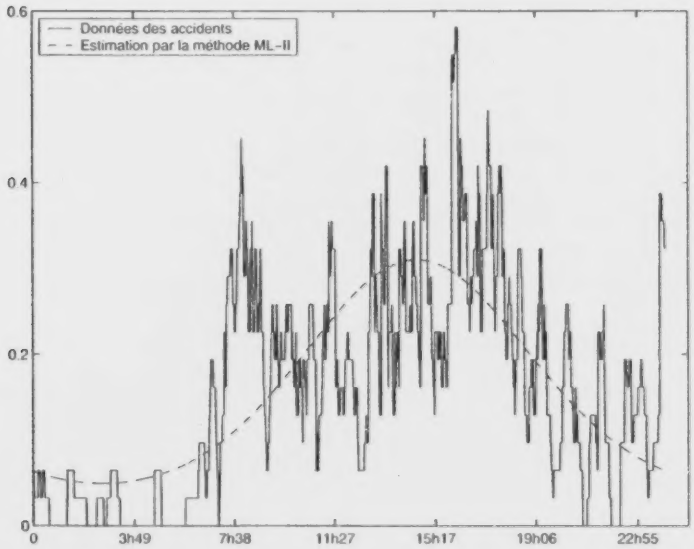

FIGURE 3.2.3. Comparaison de la répartition des accidents selon l'heure de la journée et son approximation par la méthode ML-II.

le tableau 3.2.8). Nous remarquons que le modèle à quatre modes est celui qui maximise l'entropie avec une valeur de -1,625. Nous le conservons donc comme modèle pour décrire la répartition des accidents de voiture dans une journée

$$f_{\text{acc}}(Y \mid \underset{\sim}{\theta}) = \sum_{j=1}^{4} p_j m_j(Y \mid \underset{\sim}{\theta}_j),$$

où $m_j(Y \mid \underset{\sim}{\theta}_j)$ est donné à la proposition 2.1.1 et $\underset{\sim}{\theta}_j = (\mu_j, \omega_j, \kappa)$ sont donnés au tableau 3.2.7.

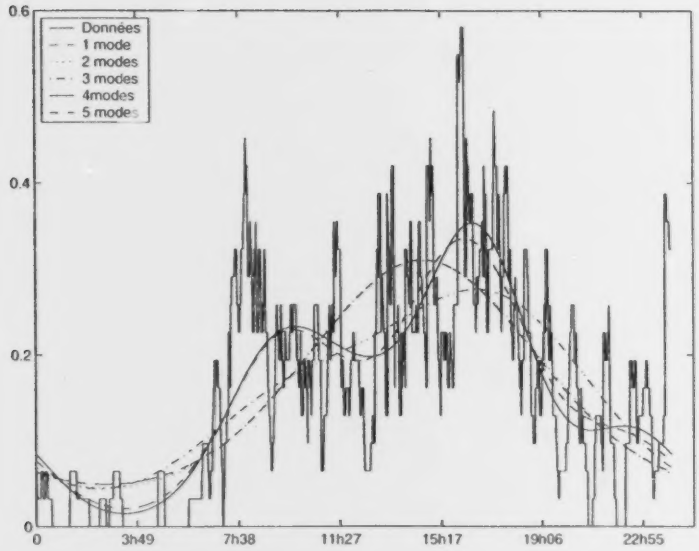

FIGURE 3.2.4. Comparaison de la répartition des accidents selon l'heure de la journée et son approximation selon le nombre de modes.

3.2.3. Probabilités marginales

Le but de cette section est ultimement de parvenir à modéliser les densités $\Upsilon 1$ et $\Upsilon 2$ des probabilités marginales du tableau de contingence 3.2.1. La densité $\Upsilon 1$ est en fait la densité reliée à la probabilité d'utiliser son téléphone mobile au moment où un accident de la route survient, alors que $\Upsilon 2$ est plutôt la densité de la probabilité d'avoir un accident de voiture au même moment. La i^e réalisation de $\Upsilon 1$ sera notée $v1_i$ alors que la i^e réalisation de $\Upsilon 2$ sera notée $v2_i \quad \forall i$.

Définition 3.2.1. *Soit T une variable aléatoire représentant le temps où un certain événement se produit. Posons $(A_{i1}\ ,\ A_{i2}), \quad i = 1, \ldots, n_{acc}$ le couple représentant l'intervalle de temps (en radian) dans lequel le i^e accident s'est produit. Pour un i fixé, la probabilité que T appartienne à cet intervalle est donnée par*

$$\mathbb{P}(T \in (A_{i1}\ ,\ A_{i2})) = \int_{A_{i1}}^{A_{i2}} f_T(t)\, dt,$$

où $f_T(t)$ est la densité de la variable aléatoire T.

TABLEAU 3.2.6. Estimation des paramètres d'un mélange de densités marginales à deux ou trois composantes pour les données sur les accidents

	Deux composantes		Trois composantes	
	Valeur initiale	**Valeur finale**	**Valeur initiale**	**Valeur finale**
p_1	0,560	0,561	0,270	0,292
p_2	0,440	0,439	0,240	0,249
p_3			0,490	0,459
μ_1	3,800 (14h31)	3,751 (14h20)	2,460 (9h24)	2,464 (9h25)
μ_2	4,100 (15h40)	4,044 (15h27)	4,010 (15h19)	4,008 (15h19)
μ_3			4,900 (21h32)	4,797 (20h45)
ω_1	1,120	1,148	280,000	280,000
ω_2	1,090	1,090	280,000	280,000
ω_3			1,000	0,962
κ	1,500	1,575	2,400	2,490

Nous pouvons donc trouver, pour chacun des n_{acc} accidents, la probabilité qu'un appel ou un accident se soit produit dans le même intervalle de temps. En effet pour l'accident i, nous n'avons qu'à remplacer la densité $f_T(t)$ par $f_{\text{app}}(x \mid \underset{\sim}{\theta})$ dans la définition 3.2.1 pour obtenir la probabilité qu'un appel se soit fait dans le même intervalle de temps. De la même façon, nous pouvons remplacer $f_T(t)$ par $f_{\text{acc}}(y \mid \underset{\sim}{\theta})$ pour trouver la probabilité qu'un accident de voiture se produise dans l'intervalle de temps $(A_{i1}\ ,\ A_{i2})$. Nous sommes donc en mesure de trouver les n_{acc} probabilités pour chacune des densités considérées et de produire les histogrammes résultants. Nous retrouvons l'histogramme des probabilités liées aux appels et celui des probabilités concernant les accidents aux figures 3.2.5 et 3.2.6 respectivement.

TABLEAU 3.2.7. Estimation des paramètres d'un mélange de densités marginales à quatre ou cinq composantes pour les données sur les accidents

	Quatre composantes		**Cinq composantes**	
	Valeur initiale	**Valeur finale**	**Valeur initiale**	**Valeur finale**
p_1	0,140	0,145	0,170	0,176
p_2	0,280	0,286	0,150	0,145
p_3	0,400	0,418	0,240	0,236
p_4	0,180	0,151	0,290	0,294
p_5			0,150	0,149
μ_1	2,300 (8h47)	2,312 (8h50)	2,400 (9h10)	2,413 (9h13)
μ_2	3,000 (11h28)	3,026 (11h34)	2,800 (10h42)	2,718 (10h23)
μ_3	4,400 (16h48)	4,382 (16h44)	3,900 (14h54)	3,977 (15h11)
μ_4	6,000 (22h55)	5,928 (22h39)	4,500 (17h11)	4,424 (16h54)
μ_5			5,600 (21h23)	5,685 (21h43)
ω_1	127,000	133,350	180,000	218,791
ω_2	1,000	0,950	5,000	5,000
ω_3	90,000	90,000	5,000	2,448
ω_4	130,000	136,500	21,000	25,526
ω_5			5,000	5,000
κ	3,500	3,763	2,000	2,549

En résumé, pour obtenir chacun des histogrammes, nous avons procédé de la façon suivante. Pour chacune des entrées de la base de données des accidents fournies par la SAAQ, nous avons considéré une fenêtre de 15 minutes précédent

TABLEAU 3.2.8. Comparaison des différents modèles pour les données sur les accidents

Modèle	$-\widehat{H}_f(\underline{\theta})$	AIC	BIC
1 mode	-1,677	-4,677	-12,466
2 modes	-1,675	-7,675	-23,233
3 modes	-1,645	-10,645	-34,012
4 modes	-1,625	-13,625	-44,781
5 modes	-1,638	-16,638	-55,584

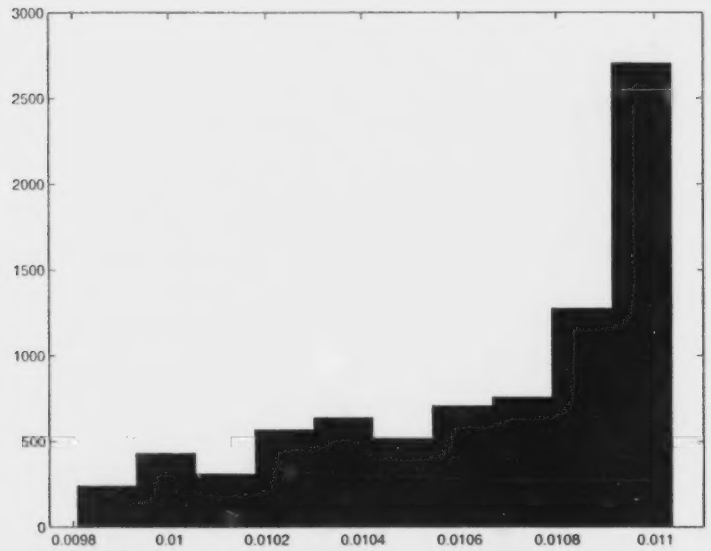

FIGURE 3.2.5. Répartition des probabilités d'appel sur un intervalle de 15 minutes.

le moment de chacun des accidents. Nous avons par la suite intégré, sur chacun des intervalles, la densité associée à la répartition des appels (voir figure 3.2.2) ou des accidents (voir figure 3.2.4) dépendant de l'histogramme que nous voulions bâtir. Par cette intégration, nous obtenons une probabilité associée à chacun des accidents. Après avoir effectué les n_{acc} intégrations, nous nous retrouvons avec n_{acc} probabilités qui nous servent à construire les histogrammes.

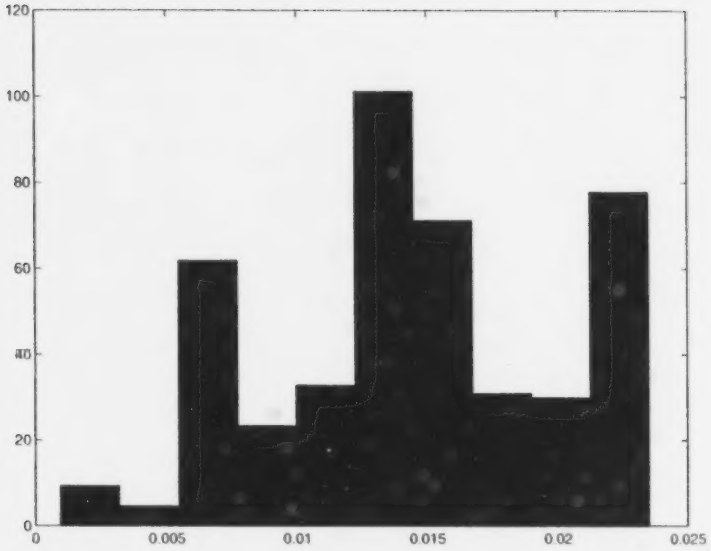

FIGURE 3.2.6. Répartition des probabilités d'accident sur un intervalle de 15 minutes.

3.2.3.1. *Modélisation par une densité bêta modifiée*

Nous allons maintenant tenter d'ajuster une courbe à chacun des histogrammes qui nous permettra de paramétriser la répartition des probabilités d'avoir un accident de voiture ou de faire l'usage du téléphone mobile au moment où se produit un accident de voiture. Comme nous voulons modéliser des probabilités, nous allons poser comme hypothèse que ces probabilités proviennent d'une densité bêta(α, β).

Définition 3.2.2. *Une variable aléatoire V est de densité bêta(α, β) si*

$$f(v \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \quad v^{\alpha-1} \quad (1-v)^{\beta-1} \quad \mathbb{I}_{[0,1]}(v)$$

où $\alpha,\ \beta > 0$. De plus,

$$\mathbb{E}_{\alpha,\beta}(V) = \frac{\alpha}{\alpha+\beta} \quad et \quad Var_{\alpha,\beta}(V) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

La pertinence du choix de la densité bêta pour l'ajustement réside dans le fait que cette densité possède le même domaine qu'une probabilité, c'est-à-dire l'intervalle [0 , 1]. Nous devons maintenant trouver les estimateurs par la méthode du maximum de vraisemblance de la densité conjointe des n_{acc} observations. Cette approche est tirée de Beckman et Tietjen (1978)

Proposition 3.2.1. *Soit* $V_i, \quad i = 1, \ldots, n_{acc}$, n_{acc} *variables aléatoires indépendantes de densité bêta*$(\alpha\ ,\ \beta)$. *Les estimateurs du maximum de vraisemblance* $\hat{\alpha}$ *et* $\hat{\beta}$ *pour* α *et* β *sont solution du système d'équations*

$$\frac{1}{n_{acc}} \sum_{i=1}^{n_{acc}} \log(v_i) = \Psi(\hat{\alpha}) - \Psi(\hat{\alpha} + \hat{\beta}),$$

$$\frac{1}{n_{acc}} \sum_{i=1}^{n_{acc}} \log(1 - v_i) = \Psi(\hat{\beta}) - \Psi(\hat{\alpha} + \hat{\beta}),$$

où $\Psi(\cdot) = \dfrac{\Gamma'(\cdot)}{\Gamma(\cdot)}$ *représente la fonction digamma.*

DÉMONSTRATION. Commençons par déterminer la densité conjointe des n_{acc} observations,

$$\begin{aligned} f(\underline{v} \mid \alpha, \beta) &= \prod_{i=1}^{n_{acc}} f(\underline{v}_i \mid \alpha, \beta) \\ &= \prod_{i=1}^{n_{acc}} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \quad v_i^{\alpha-1} \quad (1 - v_i)^{\beta-1} \\ &= \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^{n_{acc}} \quad \left[\prod_{i=1}^{n_{acc}} v_i\right]^{\alpha-1} \quad \left[\prod_{i=1}^{n_{acc}} (1 - v_i)\right]^{\beta-1}. \end{aligned}$$

En prenant le logarithme, nous obtenons

$$\begin{aligned} \log f(\underline{v} \mid \alpha, \beta) = n_{acc}[&\log\Gamma(\alpha+\beta) - \log\Gamma(\alpha) - \log\Gamma(\beta)] \\ &+ (\alpha - 1) \sum_{i=1}^{n_{acc}} \log v_i + (\beta - 1) \sum_{i=1}^{n_{acc}} \log(1 - v_i). \end{aligned} \tag{3.2.1}$$

Il ne nous reste qu'à maximiser (3.2.1) par rapport à chacun des paramètres impliqués. Donc pour α, nous obtenons

$$\frac{\partial \log f(\underset{\sim}{v} \mid \alpha, \beta)}{\partial \alpha} = n_{\text{acc}} \left[\frac{\Gamma'(\alpha+\beta)}{\Gamma(\alpha+\beta)} - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} \right] + \sum_{i=1}^{n_{\text{acc}}} \log v_i$$

$$= 0 \Leftrightarrow \frac{1}{n_{\text{acc}}} \sum_{i=1}^{n_{\text{acc}}} \log v_i = \Psi(\alpha) - \Psi(\alpha+\beta),$$

où $\Psi(\cdot) = \dfrac{\Gamma'(\cdot)}{\Gamma(\cdot)}$; et pour β,

$$\frac{\partial \log f(\underset{\sim}{v} \mid \alpha, \beta)}{\partial \beta} = n_{\text{acc}} \left[\frac{\Gamma'(\alpha+\beta)}{\Gamma(\alpha+\beta)} - \frac{\Gamma'(\beta)}{\Gamma(\beta)} \right] + \sum_{i=1}^{n_{\text{acc}}} \log(1-v_i)$$

$$= 0 \Leftrightarrow \frac{1}{n} \sum_{i=1}^{n_{\text{acc}}} \log(1-v_i) = \Psi(\beta) - \Psi(\alpha+\beta).$$

□

Afin de résoudre ce système d'équations, nous allons employer la méthode itérative de la plus grande pente décrite au chapitre 11 de Burden and Faires (1997). (Pour utiliser cet algorithme, nous avons besoin d'approximer la fonction $\Psi(a)$ pour de grandes valeurs de a. Cette approximation est tirée de Abramowitz et Stegun(1964).) Toutefois, nous devons fournir une estimation initiale des paramètres, soient $\hat{\alpha}_0$ et $\hat{\beta}_0$. Une des solutions généralement utilisées est d'avoir recours aux estimateurs donnés par la méthode des moments. Cette méthode consiste essentiellement à égaliser les moments théoriques aux moments échantillonnaux correspondants et de solutionner le système d'équations ainsi obtenu.

Proposition 3.2.2. *Soit V_i, $i = 1, \ldots, n_{acc}$ n_{acc} variables aléatoires i.i.d de densité bêta(α, β). Les estimateurs $\tilde{\alpha}$ et $\tilde{\beta}$ pour α et β obtenus par la méthode des moments sont données par*

$$\tilde{\alpha} = \bar{v}\left(\frac{\bar{v}(1-\bar{v})}{s_v^2} - 1 \right)^+ \quad et \quad \tilde{\beta} = (1-\bar{v})\left(\frac{\bar{v}(1-\bar{v})}{s_v^2} - 1 \right)^+,$$

où $\bar{v} = n_{acc}^{-1} \sum_{i=1}^{n_{acc}} v_i$ *est le premier moment échantillonnal,*

$$s_v^2 = n_{acc}^{-1} \sum_{i=1}^{n_{acc}} (v_i - \bar{v})^2 \text{ et } (a)^+ = \begin{cases} a & si \quad a > 0, \\ 0 & sinon. \end{cases}$$

DÉMONSTRATION. Nous savons que si $V \sim$ bêta(α, β) alors le premier moment théorique est simplement l'espérance de la densité donc $\mathbb{E}(V)_{\alpha,\beta} = \dfrac{\alpha}{\alpha+\beta}$. Pour cette même variable aléatoire, le k^e moment échantillonnal est donné par $M_k = n_{acc}^{-1} \sum_{i=1}^{n_{acc}} v_i^k$, $k \in \mathbb{N}$. Il nous faut donc débuter par égaliser le premier moment échantillonnal au premier moment théorique, c'est-à-dire :

$$\bar{v} = \frac{\alpha}{\alpha+\beta} \Leftrightarrow \alpha = \left(\frac{\bar{v}}{1-\bar{v}}\right)\beta. \tag{3.2.2}$$

Nous pouvons remarquer que

$$Var_{\alpha,\beta}(V) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{\bar{v}(1-\bar{v})}{(\alpha+\beta+1)}. \tag{3.2.3}$$

En remplaçant la valeur de α obtenue par l'équation (3.2.2) dans l'équation (3.2.3), nous trouvons

$$\tilde{\beta} = (1-\bar{v})\left(\frac{\bar{v}(1-\bar{v})}{s_v^2} - 1\right),$$

et en substituant cette nouvelle valeur de β dans l'équation (3.2.2), nous obtenons

$$\tilde{\alpha} = \bar{v}\left(\frac{\bar{v}(1-\bar{v})}{s_v^2} - 1\right).$$

Comme α et β sont non négatifs, nous devons uniquement considérer les parties positives de ces deux dernières équations.

□

Nous pouvons remarquer sur les histogrammes des figures 3.2.5 et 3.2.6 que les probabilités non nulles sur l'abscisse sont très petites. Ceci a pour conséquence de rendre numériquement impossible le calcul des estimateurs en Matlab (version 6.0.0.88) par la méthode du maximum de vraisemblance car le paramètre β devient très grand et aucune évaluation n'est possible pour $\Gamma(\beta)$. Pour pallier à ce problème, nous avons appliquer la transformation

$$\frac{V - D_{(1)}}{D_{(n)} - D_{(1)}}, \tag{3.2.4}$$

TABLEAU 3.2.9. Estimation des paramètres par le maximum de vraisemblance pour la distribution des probabilités des accidents et des appels

	Appels		Accidents	
	Valeur initiale	**Valeur finale**	**Valeur initiale**	**Valeur finale**
$\hat{\alpha}$	2,700	2,687	1,814	1,773
$\hat{\beta}$	1,089	1,098	1,090	1,186

où $D_{(1)}$ et $D_{(n)}$ sont respectivement la plus petite et la plus grande probabilité possible de l'événement considéré sur un intervalle de 15 minutes continues quelconque dans une journée. Cette transformation a pour conséquence d'élargir le domaine des probabilités non nulles à l'intervalle [0,1].

Nous avons donc pu trouver les estimateurs $\hat{\alpha}$ et $\hat{\beta}$ par la méthode du maximum de vraisemblace afin d'ajuster une densité bêta à nos probabilités transformées, que nous notons v pour chacun de nos histogrammes. Nous avons utilisé comme estimations initiales $\hat{\alpha}_0$ et $\hat{\beta}_0$, pour l'algorithme de la plus grande pente, les estimations résultant de la méthode des moments. Nous fournissons le tableau 3.2.9 qui indique, pour les données sur les appels et les données sur les accidents, les valeurs de départ des paramètres ainsi que les valeurs finales. Nous avons ajouter les figures 3.2.7 pour les appels et 3.2.8 pour les accidents qui permettent de comparer l'histogramme des $v_i^{\text{transf.}} = (v_i - D_{(1)})/(D_{(n)} - D_{(1)}), \quad i = 1, \ldots, n_{\text{acc}}$ avec la densité bêta obtenue pour l'ajustement. Il est à noter que les valeurs de $D_{(1)}$ et $D_{(n)}$ sont respectivement $9{,}8 \times 10^{-3}$ et 0,011 pour les appels. Quant aux accidents, les valeurs de $D_{(1)}$ et $D_{(n)}$ deviennent $9{,}198 \times 10^{-4}$ et 0,0235.

3.2.3.2. *Application de l'algorithme EM à un mélange de densités bêta*

Encore une fois, nous pouvons remarquer que l'histogramme présenté à la figure 3.2.6 montre plusieurs modes distincts. Nous avons proposé à la section 2.2.2, l'utilisation de l'algorithme itératif Espérance-Maximisation pour ajuster

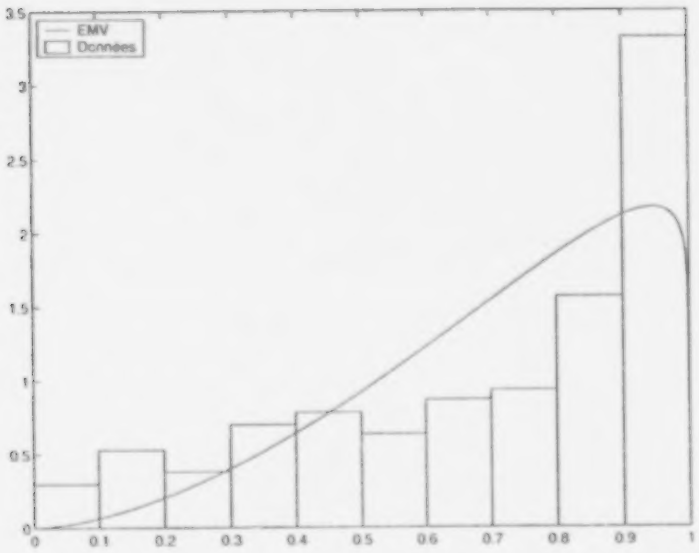

FIGURE 3.2.7. Comparaison de la répartition des probabilités transformées d'appel et son approximation par le maximum de vraisemblance

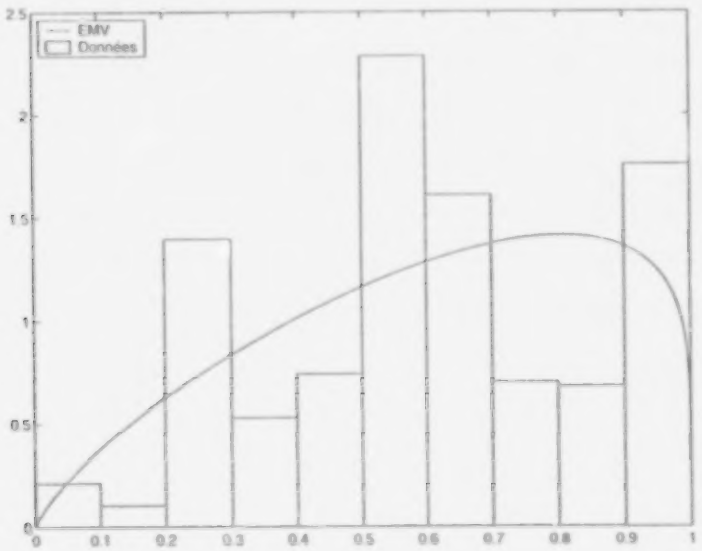

FIGURE 3.2.8. Comparaison de la répartition des probabilités transformées d'accident et son approximation par le maximum de vraisemblance

un meilleur modèle à des données multimodales. Rappelons que l'étape E consiste

à calculer

$$Q(\underset{\sim}{\psi}_g\ ,\ \underset{\sim}{\psi}_g^{(k)}) = \sum_{j=1}^{g}\sum_{i=1}^{n} \mathbb{E}^{z\ |\ \underset{\sim}{\psi}_g^{(k)}\ ,\ \underset{\sim}{y}}[Z_{ij}] \log f_j(y_i\ |\ \underset{\sim}{\theta}_j) + \sum_{j=1}^{g} \log p_j \sum_{i=1}^{n} \mathbb{E}^{z\ |\ \underset{\sim}{\psi}_g^{(k)}\ ,\ \underset{\sim}{y}}[Z_{ij}],$$

où $\mathbb{E}^{z\ |\ \underset{\sim}{\psi}_g^{(k)}\ ,\ \underset{\sim}{y}}[Z_{ij}] = \mathbb{E}^{z_{ij}\ |\ \underset{\sim}{\psi}_g^{(k)}\ ,\ \underset{\sim}{y}}[Z_{ij}] = \frac{p_j^{(k)} f_j(x_i\ |\ \underset{\sim}{\theta}_j^{(k)})}{\sum_{r=1}^{g} p_r^{(k)} f_r(x_i\ |\ \underset{\sim}{\theta}_r^{(k)})}$ et $\sum_{j=1}^{g} p_j = 1$.

Dans le cas particulier auquel nous nous intéressons, $f_j(y_i\ |\ \underset{\sim}{\theta}_j)$ est de densité bêta telle que donnée à la définition 3.2.2 avec comme vecteur de paramètres $\underset{\sim}{\theta}_j = (\alpha_j, \beta_j)$.

À l'étape M, pour un modèle à g composantes, nous allons maximiser les probabilités $p_j, \quad j = 1, \ldots, g$ à l'aide du résultat fournit par le théorème 2.2.3, c'est-à-dire pour j fixé à la k^e étape, nous obtenons

$$p_j^{(k+1)} = \frac{\sum_{i=1}^{n} \mathbb{E}^{z_{ij}\ |\ \underset{\sim}{\psi}_g^{(k)}\ ,\ \underset{\sim}{y}}[Z_{ij}]}{n},$$

où $\mathbb{E}^{z_{ij}\ |\ \underset{\sim}{\psi}_g^{(k)}\ ,\ \underset{\sim}{y}}[Z_{ij}]$ est donnée à l'équation (2.2.1). La maximisation des paramètres α_j et β_j se fait par la méthode itérative de la plus grande pente. Pour obtenir les approximations initiales, nous avons choisi arbitrairement $g-1$ limites, notée $L_1, \ldots, L_{g-1}$, sur l'abscisse qui nous semblait départager les g densités d'un mélange. Nous avons déterminé la probabilité initiale $p_{j-1}^{(0)}, \quad j = 2, \ldots, g$, simplement par $p_{j-1}^{(0)} = \text{card}(L_{j-2} < v \leq L_{j-1})/n_{acc}$, avec $L_0 = 0$ et $p_g = 1 - \sum_{j=1}^{g-1} p_j$. Pour obtenir les estimateurs initiaux $\alpha_j^{(0)}$ et $\beta_j^{(0)}$, nous avons appliqué la méthode des moments décrite précédemment à chacun des sous-échantillons définis par les limites $L_j, \quad j = 1, \ldots, g-1$. Nous fournissons les résultats des estimateurs en fonction du nombre de composantes g au tableau 3.2.10 pour les données sur les appels et au tableau 3.2.11 pour les données sur les accidents. Nous avons aussi ajouté les graphiques comparatifs 3.2.9 et 3.2.10 pour les différents modèles selon le jeu de données considéré.

Pour le choix du modèle à conserver, nous avons utilisé les mêmes critères que précédemment, c'est-à-dire le maximum d'entropie, AIC et BIC. Les résultats des différents critères selon le nombre de composantes sont présentés au tableau 3.2.12. Pour les données sur les appels ainsi que pour les données sur les accidents,

TABLEAU 3.2.10. Estimation des paramètres d'un mélange de densités bêta à deux ou trois composantes pour les données sur les appels

	Deux composantes		Trois composantes	
	Valeur initiale	**Valeur finale**	**Valeur initiale**	**Valeur finale**
p_1	0,419	0,429	0,121	0,097
p_2	0,581	0,571	0,298	0,245
p_3			0,581	0,608
α_1	2,178	1,695	2,406	2,420
α_2	9,730	9,709	9,439	5,340
α_3			9,730	9,707
β_1	3,226	2,256	13,864	13,865
β_2	1,064	1,419	9,209	5,828
β_3			1,064	1,461

TABLEAU 3.2.11. Estimation des paramètres d'un mélange de densités bêta à deux ou trois composantes pour les données sur les accidents

	Deux composantes		Trois composantes	
	Valeur initiale	**Valeur finale**	**Valeur initiale**	**Valeur finale**
p_1	0,687	0,738	0,298	0,410
p_2	0,313	0,262	0,389	0,283
p_3			0,313	0,307
α_1	4,012	2,779	4,679	2,402
α_2	8,518	8,507	73,362	61,317
α_3			8,518	9,854
β_1	4,597	2,903	10,496	3,744
β_2	1,036	1,226	51,649	42,933
β_3			1,036	1,456

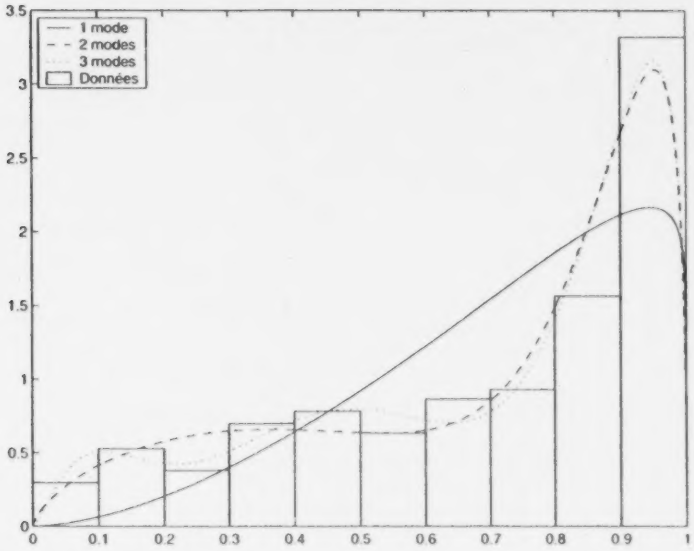

FIGURE 3.2.9. Comparaison de la répartition des probabilités transformées d'appel et son approximation par différents nombres de composantes dans un mélange de densités bêta

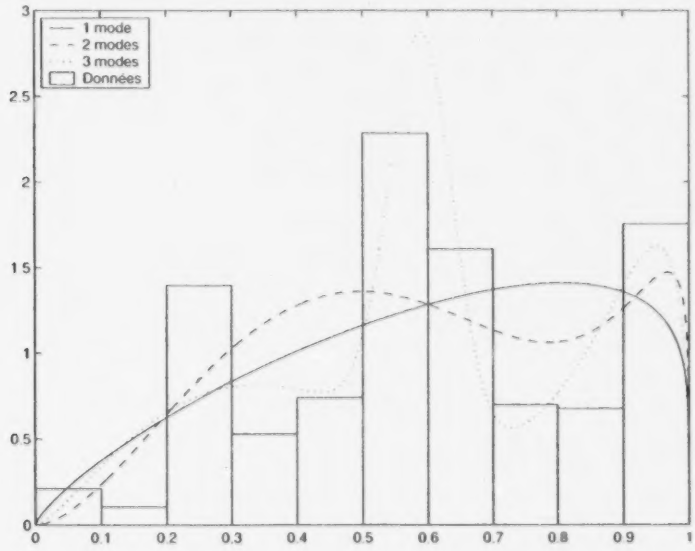

FIGURE 3.2.10. Comparaison de la répartition des probabilités transformées d'accident et son approximation par différents nombres de composantes dans un mélange de densités bêta

TABLEAU 3.2.12. Comparaison des différents modèles pour les accidents et les appels

	Appels			Accidents		
Modèle	$-\widehat{H}_f(\underline{\theta})$	**AIC**	**BIC**	$-\widehat{H}_f(\underline{\theta})$	**AIC**	**BIC**
1 mode	1,526	-0,474	-2,795	1,316	-0,684	-3,006
2 modes	2,993	-2,007	-7,810	2,964	-2,036	-7,839
3 modes	6,230	-1,770	-11,055	3,090	-1,910	-7,713

le modèle à conserver est celui avec trois composantes de densité bêta car ils maximisent le critère du maximum d'entropie. En somme, les modèles considérés pour la modélisation des densités des probabilités marginales sont :

Pour les appels : $\Upsilon 1^{\text{transf.}} = \sum_{j=1}^{3} p_j f_{j_{\text{app}}}(\upsilon 1^{\text{transf.}} \mid \underline{\theta}_j)$, où les $\underline{\theta}_j = (\alpha_j,\ \beta_j)$ sont données au tableau 3.2.10.

Pour les accidents : $\Upsilon 2^{\text{transf.}} = \sum_{j=1}^{3} p_j f_{j_{\text{acc}}}(\upsilon 2^{\text{transf.}} \mid \underline{\theta}_j)$, où les $\underline{\theta}_j = (\alpha_j,\ \beta_j)$ sont données au tableau 3.2.11.

Nous possédons maintenant les densités associées aux probabilités marginales $\Upsilon 1$ (être au téléphone) et $\Upsilon 2$ (avoir un accident) à une des heures d'accidents observées dans notre base de données. Nous allons voir dans la prochaine section comment utiliser ces deux densités dans un tableau de contingence afin de calculer une approximation du risque relatif ou du rapport de cotes.

3.3. RECONSTRUCTION DU TABLEAU DE CONTINGENCE

Afin de calculer des mesures d'association entre les variables aléatoires X, "être au téléphone mobile en conduisant" et Y, "avoir un accident de voiture", nous avons proposé à la section 3.2 de construire un tableau de contingence 2×2 du type suivant :

Grâce aux données recueillies pour l'ensemble des participants étant les seuls utilisateurs du téléphone mobile, nous avons de l'information sur l'utilisation du téléphone mobile ainsi que sur le dossier de conduite (pour les accidents) pour

X / Y	Oui	Non	Total
Oui	n_{11}		$\Upsilon 1$
Non	n_{21}		
Total	$\Upsilon 2$		n

6360 d'entre eux. Toutefois, lorsque nous considérions un appel de t minutes dans la modélisation de $f_{\text{app}}(X \mid \underset{\sim}{\mu}, \underset{\sim}{\omega}, \underset{\sim}{\kappa})$, cet appel comptait véritablement comme t appels/minute. De même, pour la modélisation de $f_{\text{acc}}(Y \mid \underset{\sim}{\mu}, \underset{\sim}{\omega}, \underset{\sim}{\kappa})$, chaque accident était considéré comme 15 accidents/minute (par l'hypothèse d'une fenêtre de temps de 15 minutes). Afin de trouver le nombre n d'individus/minute qui ont pris part à l'étude, nous devons considérer deux quantités : le nombre de personnes dans l'étude et le nombre de personnes impliquées dans plus d'un accident. Dans le premier cas, nous avons 6360 personnes dans l'étude. Nous pouvons également remarquer qu'un total de 473 accidents ont été faits par 442 personnes différentes. Donc, 31 personnes vont compter pour plus d'un individu/minute. Par conséquent, comme nous avons considéré un intervalle de 15 minutes pour les accidents, le nombre d'individus/minute pour le groupe des seuls utilisateurs devient donc $n = 15 * (6360 + 31) = 95865$.

Nous devons maintenant établir la valeur de n_{11}, le nombre d'individus/minute qui faisaient usage de leur téléphone mobile dans un intervalle de temps de 15 minutes précédant le moment où ils ont eu un accident de la route. Il faut toutefois porter attention à certaines situations. Un individu a pu faire usage de son téléphone plusieurs fois durant cet intervalle de temps. Nous avons donc considéré la durée du dernier appel fait ou reçu avant le moment de l'accident. Un individu a aussi pu débuter une conversation téléphonique avant le début de l'intervalle de 15 minutes. Dans ce cas, nous avons uniquement comptabilisé la durée de l'appel fait dans l'intervalle. Enfin, un individu a pu mettre fin à un appel après le moment défini par la borne supérieure de l'intervalle. Par exemple, si quelqu'un faisait usage de son téléphone au moment de l'accident, il est peu probable que l'appel se termine au moment précis de l'accident. Dans ce dernier cas, nous avons comptabilisé uniquement la durée de l'appel jusqu'au moment défini par la borne

supérieure de l'intervalle. Nous pouvons donc trouver n_{11} qui est la somme de ces durées d'appels faits dans un intervalle de 15 minutes où les individus ont eu un accident de la route. Pour le cas des seuls utilisateurs avec tous types d'accidents, nous trouvons $n_{11} = 140$ accidents/minute. Comme nous avions trouvé qu'un total de 473 accidents étaients survenus, c'est donc dire que $n_{+1} = 7095$ et donc $n_{21} = 7095 - 140 = 6955$.

Pour le calcul de n_{11}, dans une situation où un individu a fait usage de son téléphone mobile plusieurs fois dans l'intervalle de 15 minutes, nous avons comptabilisé le nombre de minutes du dernier appel. Ce choix peut être discutable. Toutefois nous avons aussi déterminé, chaque fois où cette situation survenait, le temps minimum des appels et le temps maximum des appels dans l'intervalle de 15 minutes afin d'obtenir en quelques sortes des bornes inférieure et supérieure pour n_{11}. En comptabilisant les temps minimum des appels, nous obtenons $n_{11}^{\min} = 111$ et avec les temps maximum, nous trouvons $n_{11}^{\max} = 184$. Nous pouvons remarquer que la valeur de n_{11} trouvée est un temps intermédiaire entre les bornes définies par $n_{11}^{\min}$ et $n_{11}^{\max}$. Nous allons donc pouvoir utiliser ces quantités dans l'application de la méthode de Monte Carlo avec fonction d'importance.

3.3.1. Méthode de Monte Carlo avec fonction d'importance

Avec les valeurs maintenant connues de n_{11}, n_{21}, n et des densités $\Upsilon 1$ et $\Upsilon 2$, nous sommes en mesure d'estimer le rapport de cotes (RC) et le risque relatif (RR) pour les variables aléatoires catégorielles "avoir un accident de la route" et "faire usage de son téléphone mobile en conduisant". Pour y arriver, nous allons utiliser la méthode de Monte Carlo avec fonction d'importance (voir Robert, 1994, section 9.2.2).

Dans un problème statistique bayésien, nous avons souvent à calculer

$$\mathbb{E}[g(\theta)] = \int_{\Theta} g(\theta) f(\underset{\sim}{x}|\theta)\pi(\theta)d\theta. \tag{3.3.1}$$

Il arrive que cette intégrale soit difficile à évaluer, alors dans un tel cas nous pouvons approximer cette intégrale par une moyenne.

Définition 3.3.1. *Si h est une densité quelconque sur l'espace des paramètres Θ, nous définissons la méthode de Monte Carlo avec fonction d'importance de la façon suivante : générer $\theta_1, \ldots, \theta_r$, r vecteurs de paramètres à partir de h et approximer l'équation* (3.3.1) *par*

$$\frac{1}{r}\sum_{i=1}^{r} g(\theta_i)w_i(\theta_i),$$

avec des poids $w(\theta_i) = f(\underset{\sim}{x}|\theta_i)\pi(\theta_i)/h(\theta_i)$.

Par conséquent pour une espérance conditionnelle, nous obtenons

$$\begin{aligned}\mathbb{E}[g(\theta) \mid \underset{\sim}{x}] &= \frac{\int_\Theta g(\theta|\underset{\sim}{x}) f(\underset{\sim}{x}|\theta)\pi(\theta)d\theta}{\int_\Theta f(\underset{\sim}{x}|\theta)\pi(\theta)d\theta} \\ &\approx \frac{1/r\sum_{i=1}^{r} g(\theta_i)w(\theta_i)}{1/r\sum_{i=1}^{r} w(\theta_i)}. \end{aligned} \tag{3.3.2}$$

Définition 3.3.2. *La variance de l'approximation définie à l'équation* (3.3.2) *par la méthode de Monte Carlo avec fonction d'importance est donnée par*

$$\frac{S_N^2 - 2(\bar{N}/\bar{D})S_{ND} + (\bar{N}/\bar{D})^2 S_D^2}{r(r-1)(\bar{D})^2}, \tag{3.3.3}$$

où $\bar{N}$ représente le numérateur de l'équation (3.3.2), *$\bar{D}$, son dénominateur et*

$S_N^2 = \sum_{i=1}^{r}(g(\theta_i)w(\theta_i) - \bar{N})^2$,

$S_D^2 = \sum_{i=1}^{r}(w(\theta_i) - \bar{D})^2$,

$S_{ND} = \sum_{i=1}^{r}[(g(\theta_i)w(\theta_i) - \bar{N})(w(\theta_i) - \bar{D})]$.

La précision de l'approximation donnée à l'équation (3.3.2) *correspond à la racine carrée de l'équation* (3.3.3).

Afin d'appliquer cette méthode, nous devons connaître la densité conjointe des observations ainsi que chacune des densités *a priori* des différents paramètres. Nous avons constaté à la section 1.1.1 que la répartition de n épreuves dans un tableau de contingence 2×2 suit une loi multinomiale de paramètres $\underset{\sim}{n}$, $\underset{\sim}{\nu}$ où $\underset{\sim}{n} = (n_{11}, n_{12}, n_{21}, n_{22})$ et $\underset{\sim}{\nu} = (\nu_{11}, \nu_{12}, \nu_{21}, \nu_{22})$. Nous pouvons donc écrire la

fonction de vraisemblance des observations comme

$$L(\nu_{11}, \nu_{12}, \nu_{21} \mid \underset{\sim}{n}) = L(\nu_{12} \mid \nu_{11}, \nu_{21}, \underset{\sim}{n}) L(\nu_{11}, \nu_{21} \mid \underset{\sim}{n}).$$

Or, $L(\nu_{11}, \nu_{21} \mid \underset{\sim}{n})$ peut être vu comme une multinomiale avec uniquement trois catégories possibles. Donc, nous pouvons trouver

$$L(\nu_{11}, \nu_{21} \mid \underset{\sim}{n}) = \frac{n!}{n_{11}! n_{21}! (n - n_{11} - n_{21})!} \nu_{11}^{n_{11}} \nu_{21}^{n_{21}} (1 - \nu_{11} - \nu_{21})^{(n - n_{11} - n_{21})}.$$

Nous pouvons maintenant déduire $L(\nu_{12} \mid \nu_{11}, \nu_{21}, \underset{\sim}{n})$. En effet,

$$\begin{aligned}
L(\nu_{12} \mid \nu_{11}, \nu_{21}, \underset{\sim}{n}) &= \frac{L(\nu_{11}, \nu_{12}, \nu_{21} \mid \underset{\sim}{n})}{L(\nu_{11}, \nu_{21} \mid \underset{\sim}{n})} \\
&= \frac{\binom{n}{n_{11} n_{12} n_{21}} \nu_{11}^{n_{11}} \nu_{12}^{n_{12}} \nu_{21}^{n_{21}} (1 - \nu_{11} - \nu_{12} - \nu_{21})^{(n - n_{11} - n_{12} - n_{21})}}{\binom{n}{n_{11} n_{21}} \nu_{11}^{n_{11}} \nu_{21}^{n_{21}} (1 - \nu_{11} - \nu_{21})^{(n - n_{11} - n_{21})}} \\
&= \binom{n - n_{11} - n_{21}}{n_{12}} \frac{\nu_{12}^{n_{12}} (1 - \nu_{11} - \nu_{12} - \nu_{21})^{(n - n_{11} - n_{12} - n_{21})}}{(1 - \nu_{11} - \nu_{21})^{(n - n_{11} - n_{21})}} \\
&= \binom{n - n_{11} - n_{21}}{n_{12}} \left(\frac{\nu_{12}}{1 - \nu_{11} - \nu_{21}} \right)^{n_{12}} \\
&\qquad \times \left(\frac{1 - \nu_{11} - \nu_{12} - \nu_{21}}{1 - \nu_{11} - \nu_{21}} \right)^{(n - n_{11} - n_{12} - n_{21})} \\
&= \binom{n^*}{n_{12}} (\nu_{12}^*)^{n_{12}} (1 - \nu_{12}^*)^{n^* - n_{12}}
\end{aligned}$$

où $n^* = n - n_{11} - n_{21}$ et $\nu_{12}^* = \dfrac{\nu_{12}}{1 - (\nu_{11} + \nu_{21})} = \dfrac{v1 - \nu_{11}}{1 - v2}$ avec $v1$ et $v2$, des réalisations de $\Upsilon 1$ et $\Upsilon 2$ respectivement. C'est donc dire que

$n_{12} \mid \nu_{11},\ n_{11},\ n_{21},\ n\ ,\ v1\ ,\ v2 \sim bin(n^*\ ,\ \nu_{12}^*)$.

Nous pouvons maintenant nous intéresser à la fonction de poids. Par définition,

$$\begin{aligned}
w(\underset{\sim}{\theta_i}) &= \frac{f(\underset{\sim}{n} \mid \underset{\sim}{\theta_i}) \pi(\underset{\sim}{\theta_i})}{h(\underset{\sim}{\theta_i})} \\
&= L(\nu_{12} \mid \nu_{11}, \nu_{21}, \underset{\sim}{n}) L(\nu_{11}, \nu_{21} \mid \underset{\sim}{n}) \pi_3(\nu_{11} | v1, v2) \pi_1(v1^{\text{transf.}}) \pi_2(v2^{\text{transf.}}) / h(\underset{\sim}{\theta_i}),
\end{aligned}$$

où $\pi_1(v1^{\text{transf.}})$ et $\pi_2(v2^{\text{transf.}})$ sont les mélanges de trois densités bêta trouvées plus haut respectivement pour les appels et les accidents. Comme nous n'avons pas véritablement d'information précise pour ν_{11}, nous allons supposer $\pi_3(\nu_{11} | v1, v2)$ uniforme sur l'intervalle $[\max(0, v1 + v2 - 1)\ ,\ \min(v1, v2)]$. La borne supérieure de cet intervalle est déterminée par le fait que ν_{11} ne peut prendre une valeur

supérieure à la valeur minimum des probabilités marginales, donc $\min(v1, v2)$. Quant à la borne inférieure, elle doit satisfaire deux conditions. Comme ν_{11} est une probabilité, la borne inférieure doit être supérieure ou égale à 0. Ensuite, nous devons nous assurer que $\sum_{(i,j)} \nu_{ij} = 1, \quad i, j = 1, 2$. Ceci entraîne que ν_{11} ne peut être inférieur à $v1 + v2 - 1$. Par conséquent, la valeur de la borne inférieure est donnée par $\max(0, v1 + v2 - 1)$. Par conséquent, en posant $\nu_0 = \nu_{11}/v2$, nous obtenons $\pi_0(\nu_0 \mid v1, v2)$ qui est une densité uniforme sur l'intervalle $[\max(0\ ,\ (v1 + v2 - 1)/v2)\ ,\ \min(v1/v2\ ,\ 1)]$. De plus, la fonction de poids peut s'écrire comme

$$
\begin{aligned}
w(\underset{\sim}{\theta}_i) &= \binom{n}{n_{11} n_{21}} \nu_0^{n_{11}}\ (1 - \nu_0)^{n_{21}} \\
&\quad \times v2^{(n_{11}+n_{21}+1)}\ (1 - v2)^{(n - n_{11} - n_{21})} \\
&\quad \times L(\nu_{12} \mid \nu_{11},\ \nu_{21},\ \underset{\sim}{n})\ \pi_1(v1^{\text{transf.}})\ \pi_2(v2^{\text{transf.}})\ \pi_0(\nu_0 | v1, v2)/h(\underset{\sim}{\theta}_i) \\
&= \binom{n}{n_{11} n_{21}} \pi_4(\nu_0 \mid n_{11},\ n_{21})\ \pi_3(v2 \mid n_{11},\ n_{21}\ n) \\
&\quad \times \pi_6(n_{12} | \nu_{11},\ \nu_{21},\ n\ ,\ v1\ ,\ v2)\ \pi_1(v1^{\text{transf.}})\ \pi_2(v2^{\text{transf.}})\ \pi_0(\nu_0 \mid v1, v2)/h(\underset{\sim}{\theta}_i),
\end{aligned}
$$

Afin d'obtenir une fonction de poids plus stable, nous avons multiplié au numérateur et au dénominateur par la quantité $(v2^{\text{transf.}})^{n_{11}+n_{21}+1}(1 - v2^{\text{transf.}})^{n - n_{11} - n_{21}}$. Ceci entraîne donc

$$
\begin{aligned}
w(\underset{\sim}{\theta}_i) &= \binom{n}{n_{11} n_{21}} \pi_4(\nu_0 \mid n_{11},\ n_{21})\ \pi_5(v2 \mid n_{11},\ n_{21}\ n) / \ \pi_5(v2^{\text{transf.}} \mid n_{11},\ n_{21}\ n) \\
&\quad \times \pi_6(n_{12} | \nu_{11},\ \nu_{21},\ n\ ,\ v1\ ,\ v2)\ \pi_1(v1^{\text{transf.}})\ \pi_2^*(v2^{\text{transf.}})\ \pi_0(\nu_0 \mid v1, v2)/h(\underset{\sim}{\theta}_i),
\end{aligned}
$$

où

- $\pi_0(\nu_0 \mid v1, v2)$ est une densité uniforme sur l'intervalle $[\max(0\ ,\ (v1 + v2 - 1)/v2)\ ,\ \min(v1/v2\ ,\ 1)]$,

- $\pi_1(v1^{\text{transf.}})$ est le mélange des trois densités bêta trouvées précédemment,
- $\pi_2^*(v2^{\text{transf.}})$ devient un mélange de trois densités bêta tel que $\sum_{j=1}^{3} p_j f_{j_{\text{acc}}}(v2^{\text{transf.}} \mid \underset{\sim}{\theta}_j^*)$ où $\underset{\sim}{\theta}_j^* = (\alpha_j + n_{11} + n_{21} + 1\ ,\ \beta_j + n - n_{21})$,
- $\pi_4(\nu_0 \mid n_{11},\ n_{21})$ est une densité bêta$(n_{11} + 1\ ,\ n_{21} + 1)$,
- $\pi_5(v2 \mid n_{11},\ n_{21}\ n)$ est une densité bêta$(n_{11} + n_{21}\ ,\ n - n_{11} - n_{21} + 1)$,

TABLEAU 3.3.1. Estimation du rapport de cotes et du risque relatif par la méthode de Monte Carlo avec fonction d'importance pour n_{11}

	Estimateur	**Précision**
Risque relatif	1,735	0,001
Rapport de cotes	1,842	0,001

TABLEAU 3.3.2. Estimation du rapport de cotes et du risque relatif par la méthode de Monte Carlo avec fonction d'importance pour $n_{11}^{\min}$

	Estimateur	**Précision**
Risque relatif	1,407	0,001
Rapport de cotes	1,454	0,001

- $\pi_6(n_{12} \mid \nu_{11},\ n_{11},\ n_{21},\ n\ ,\ v1\ ,\ v2)$ est une densité $\text{bin}(n^*\ ,\ \nu_{12}^*)$.

En posant comme fonction d'importance

$$h(\underset{\sim}{\theta}_i) = \pi_1(v1^{\text{transf.}})\ \pi_2^*(v2^{\text{transf.}})\ \pi_6(n_{12} \mid \nu_{11},\ n_{11},\ n_{21},\ n\ ,\ v1\ ,\ v2)\pi_4(\nu_0 \mid n_{11},\ n_{21}),$$

nous pouvons générer des valeurs $\underset{\sim}{\theta}_i, \quad i = 1, \ldots, r$. Si nous posons dans l'équation (3.3.2), $g(\underset{\sim}{\theta} \mid n_{11},\ n_{21},\ n) = \dfrac{n_{11}n_{22}}{n_{12}n_{21}}$, où $n_{22} = n - n_{11} - n_{12} - n_{21}$ alors nous trouverons une estimation du rapport de cotes alors que si nous posons $g(\underset{\sim}{\theta} \mid n_{11},\ n_{21},\ n) = \dfrac{n_{11}(n_{21} + n_{22})}{n_{21}(n_{11} + n_{12})}$, nous obtiendrons une estimation du risque relatif. Les tableaux 3.3.1, 3.3.2 et 3.3.3, nous donnent l'estimation du rapport de cotes (RC) ainsi que celle du risque relatif (RR) obtenues à partir des données des appels et des accidents selon l'hypothèse faite pour le calcul de n_{11}. Ce tableau présente également la valeur de la précision telle que discutée à la définition 3.3.2 pour chacune des estimations.

Nous pouvons également produire des intervalles de confiance pour $g(\underset{\sim}{\theta})$ à partir des poids $w(\underset{\sim}{\theta}_i), \quad i = 1, \ldots, r$.

TABLEAU 3.3.3. Estimation du rapport de cotes et du risque relatif par la méthode de Monte Carlo avec fonction d'importance pour $n_{11}^{\max}$

	Estimateur	**Précision**
Risque relatif	2,211	0,001
Rapport de cotes	2,444	0,002

Définition 3.3.3. *Soient* G_i, $i = 1, \ldots, r$, r *variables aléatoires i.i.d. L'intervalle* $IC_{\underset{\sim}{g}} = [IC_1(\underset{\sim}{x}) \ ; \ IC_2(\underset{\sim}{x})]$ *est appelé la région* α*-crédible pour* $g(\theta)$ *avec probabilité égale dans les queues si elle respecte l'inégalité*

$$\mathbb{P}(g(\underset{\sim}{\theta}) \in IC_{\underset{\sim}{g}} \mid \underset{\sim}{g}) = 1 - \alpha.$$

et tel que $\mathbb{P}(g(\underset{\sim}{\theta}) \leq IC_1(\underset{\sim}{x})) = \mathbb{P}(g(\underset{\sim}{\theta}) \geq IC_2(\underset{\sim}{x})) = \alpha/2$.

Nous sommes donc intéressés à trouver $IC_1(\underset{\sim}{x})$ et $IC_2(\underset{\sim}{x})$. Pour ce faire, nous allons utiliser le même raisonnement que lors de la méthode de Monte Carlo avec fonction d'importance, c'est-à-dire :

$$\begin{aligned} 1 - \alpha &= \mathbb{P}(a \leq g(\underset{\sim}{\theta}) \leq b) = \int_{\Theta} \mathbb{I}_{(a,b)}(g(\underset{\sim}{\theta}))\pi(\underset{\sim}{\theta} \mid \underset{\sim}{g})d\theta \\ &\approx \frac{\sum_{i=1}^{r} \mathbb{I}_{[a,b]}(g(\underset{\sim}{\theta}_i))\ w(\underset{\sim}{\theta}_i)}{\sum_{i=1}^{r} w(\underset{\sim}{\theta}_i)}, \end{aligned}$$

où $w(\underset{\sim}{\theta}_i)$ sont les poids utilisés lors du calcul de l'approximation de $g(\underset{\sim}{\theta})$ par la méthode de Monte Carlo avec fonction d'importance. Si nous posons $\alpha = 0{,}95$ pour le groupe des seuls utilisateurs du téléphone mobile (tous types d'accidents confondus), nous obtenons les résultats du tableau 3.3.4 selon l'hypothèse faite sur le choix de n_{11}.

Sur l'ensemble des 473 accidents dans lesquels ont été impliqué des participants, 80 conducteurs ont fait usage d'un téléphone mobile dans une fenêtre de

TABLEAU 3.3.4. Comparaison des régions α-crédibles pour le risque relatif et le rapport de cotes en considérant n_{11}, $n_{11}^{\min}$ et $n_{11}^{\max}$ au niveau $1 - \alpha = 0{,}95$.

Hypothèse	**Risque relatif**	**Rapport de cotes**
n_{11}	[1,632 ; 1,841]	[1,718 ; 1,972]
$n_{11}^{\min}$	[1,323 ; 1,491]	[1,358 ; 1,552]
$n_{11}^{\max}$	[2,090 ; 2,332]	[2,286 ; 2,605]

temps de 15 minutes précédant l'heure inscrite sur le rapport de police. Pour seulement 24 accidents, le choix du temps d'appel à considérer pour le calcul de n_{11} est différent. Le mode des différences entre les temps maximum et les temps du dernier appel est de 0 tandis qu'il est de 1 pour les temps minimum. Dans les deux cas, la médiane des observations est de 1. Malgré tout, peu importe le choix de l'hypothèse sur n_{11}, les estimations du risque relatif et du rapport de cotes sont toujours supérieures à un. Donc, il y a vraisemblablement une association entre le fait de conduire un véhicule tout en utilisant un téléphone mobile et le fait d'être impliqué dans un accident de la route.

3.4. APPLICATION DE LA MÉTHODE À DES GROUPES SPÉCIFIQUES

Nous nous sommes jusqu'à présent intéressés au groupe des individus étant les seuls utilisateurs de leur téléphone mobile et pour tous types d'accidents. Nous pouvons appliquer la méthode développée à d'autres groupes cibles. Nous avons effectué l'estimation du rapport de cotes, du risque relatif ainsi que des intervalles de confiance correspondants pour trois autres groupes. Dans un premier temps, nous avons encore considéré l'ensemble des conducteurs étant les seuls utilisateurs de leur téléphone mobile mais cette fois pour les accidents ayant impliqués des blessés (mineurs, majeurs ou mortels). Les résultats des diverses étapes sont fournis à l'annexe B. Nous retrouvons les résultats finaux pour ce groupe au tableau 3.4.1.

TABLEAU 3.4.1. Estimation du rapport de cotes et du risque relatif par la méthode de Monte Carlo avec fonction d'importance pour les seuls utilisateurs pour les accidents avec blessés

	Estimateur	**Précision**
Risque relatif	2,000	0,001
Rapport de cotes	2,041	0,002

TABLEAU 3.4.2. Estimation du rapport de cotes et du risque relatif par la méthode de Monte Carlo avec fonction d'importance pour les grands utilisateurs pour tous types d'accidents

	Estimateur	**Précision**
Risque relatif	2,270	0,003
Rapport de cotes	2,532	0,004

Dans un deuxième temps, nous avons ciblé les grands utilisateurs de téléphones cellulaires (1597 individus ayant participé à l'étude ont une moyenne d'au moins sept appels par jour) et ce pour tous les types d'accidents. Enfin, le groupe des petits utilisateurs de téléphone mobile (1892 individus ont une moyenne de moins d'un appel par jour) pour tous les types d'accidents. Pour le groupe des grands utilisateurs, les résultats intermédiaires sont fournis à l'annexe C et les résultats finaux sont présentés au tableau 3.4.2. Finalement, l'annexe D contient les résultats des différentes étapes de la méthode et le tableau 3.4.3 contient les résultats finaux pour le groupe des petits utilisateurs. Nous avons ajouté le tableau 3.4.4 qui permet de comparer les intervalles de confiance des quatres groupes étudiés pour les différentes mesures d'association.

Dans ce dernier chapitre, nous avons développé une méthode qui nous a permis de mesurer le niveau d'association entre la variable "faire usage de son téléphone mobile en conduisant" et "avoir un accident de la route" par le calcul d'approximations du rapport de cotes et du risque relatif. Pour le groupe des seuls utilisateurs du téléphone mobile en considérant tous les types d'accidents, nous avons trouvé une approximation du risque relatif de 1,735 ; ce qui signifie que la probabilité

TABLEAU 3.4.3. Estimation du rapport de cotes et du risque relatif par la méthode de Monte Carlo avec fonction d'importance pour les petits utilisateurs pour tous types d'accidents

	Estimateur	**Précision**
Risque relatif	0,837	0,001
Rapport de cotes	0,832	0,001

TABLEAU 3.4.4. Comparaison des régions α-crédibles pour le risque relatif et le rapport de cotes avec l'hypothèse n_{11}, pour chacun des groupes de niveau $1 - \alpha = 0{,}95$.

Groupe	**Risque relatif**	**Rapport de cotes**
tous	[1,632 ; 1,841]	[1,718 ; 1,972]
blessés	[1,884 ; 2,136]	[1,918 ; 2,187]
grands	[2,034 ; 2,529]	[2,220 ; 2,886]
petits	[0,736 ; 0,957]	[0,730 ; 0,956]

d'avoir un accident de voiture augmente d'environ 74% si on fait usage d'un téléphone mobile tout en conduisant. Pour les usagers de téléphone mobile faisant en moyenne sept appels ou plus par jour, ce risque relatif augmente à 2,270. Tandis que ceux ayant une moyenne d'appels par jour inférieure à un ont un risque de 0,837 rendant les deux variables presqu'indépendantes. Enfin en se limitant qu'aux seuls accidents avec blessés pour l'ensemble des seuls utilisateurs, l'estimation du risque relatif se situe à 2,000. De façon générale, nous pouvons affirmer que l'utilisation du téléphone mobile tout en conduisant augmente significativement le risque d'être impliqué dans un accident de la route.

CONCLUSION

Ce mémoire nous a permis de quantifier l'association entre les variables "faire l'usage d'un téléphone mobile pendant la conduite" et "avoir un accident de voiture".

Nous avons d'abord, au premier chapitre, fait un survol de la théorie sur les tableaux de contingence 2×2 permettant de vérifier si une association quelconque existe entre deux variables aléatoires catégorielles. Nous avons ensuite présenté le rapport de cotes et le risque relatif comme deux quantités permettant de fournir une mesure du degré d'association entre ces deux variables. Nous avons également introduit brièvement la densité de von Mises utilisée pour l'analyse de données circulaires. En effet, les moments où sont survenus les accidents de même que les intervalles de temps durant lesquels un téléphone mobile est en fonction peuvent être bien modélisés par des mesures circulaires.

Au second chapitre, nous avons utilisé des éléments de statistique bayésienne afin de modéliser la répartition des appels (faits ou reçus) de même que la répartition des accidents selon le moment de la journée pour les individus étant les seuls utilisateur d'un téléphone et pour tous les types d'accidents. Dans un premier temps, nous avons utilisé la méthode du maximum de vraisemblance de type II (ML-II) pour ajuster une courbe aux données de répartition. Dans un deuxième temps, nous avons utilisé l'algorithme EM conjointement avec la méthode ML-II dans un contexte de mélange de lois afin d'obtenir une densité qui tienne compte de la multimodalité des histogrammes des données d'appels et d'accidents. Comme nous obtenions plusieurs modèles possibles qui sont fonction

du nombre de composantes dans le mélange de lois, nous devons pouvoir choisir entre les différents modèles. Le critère du maximum d'entropie a été utilisé pour sélectionner le "meilleur" modèle. Lorsque ce critère ne nous permettait pas de conclure, les critères AIC et BIC ont été utilisés. Pour les appels, un modèle unimodal a été retenu alors que pour les accidents, ce fut un mélange de lois à quatre composantes.

Le dernier chapitre proposait d'utiliser les densités obtenues pour la répartition des appels et des accidents afin de trouver, pour chacun des intervalles de temps où est survenu un accident, la probabilité qu'un téléphone mobile soit en fonction ou qu'un accident de voiture se soit produit. Nous avons pu tracer l'histogramme de ces probabilités autant pour les appels que pour les accidents. Ensuite, nous avons ajusté une densité bêta à chacun des histogrammes. Dans le but d'obtenir un meilleur ajustement, nous avons encore une fois utiliser l'algorithme EM appliqué à un mélange de lois bêta. Dans les deux cas, un modèle à trois composantes a été choisi. Ces mélanges de densités obtenus pour les appels et pour les accidents ont été utiles pour la reconstruction du tableau de contingence qui nous a servi pour le calcul de l'estimation du rapport de cotes et du risque relatif. En effet, nous avons ensuite utilisé la méthode de Monte Carlo avec fonction d'importance afin d'obtenir une estimation des deux mesures d'association, de même que des régions α-crédibles.

Globalement, nous pouvons remarquer une hausse significative du risque d'accident lorsqu'un individu fait usage d'un téléphone mobile tout en conduisant sauf pour le groupe des petits utilisateurs. En effet, le risque relatif se situe à 1,74 pour le groupe des seuls utilisateurs avec tous les types d'accidents. Ce qui veut donc dire qu'un individu utilisant son téléphone mobile au volant a environ 74% plus de chance d'être impliqué dans un accident de la route que celui ne faisant pas usage d'un téléphone mobile. Ce risque grimpe à 2,27 pour le groupe des seuls grands utilisateurs (moyenne d'au moins sept appels par jour) pour tous les types d'accidents. Si nous nous restreignons aux accidents impliquant des blessés pour

le groupe des seuls utilisateurs, le risque devient 2,00. Finalement les petits utilisateurs de téléphone mobile (pour tous les types d'accidents) ont un risque relatif qui se situe à 0,837.

Annexe A

DONNÉES UTILISÉES POUR L'EXEMPLE

2.2	5.97	3.4724	2.45	2.234
1.8481	0.43456	1.0219	0.8646	1.0904
2.67	4.06	1.7941	3.1751	1.4266
4.713	0.69971	1.9289	0.36294	6.1455
0.7215	1.0327	2.9395	1.5086	0.19676
5.0207	2.5063	4.123	5.7966	5.122
4.9905	4.8701	0.9447	0.9	0.6765
1.0788	3.7806	1.4617	5.1661	3.3924
5.972	3.6521	4.888	1.2004	5.1389
3.5836	3.9909	3.7552	4.0045	1.1355
1.9435	3.8781	4.1226	3.892	0.09756
0.4992	3.9536	4.8454	3.1454	3.9889
3.4029	5.3302	2.385	5.9166	4.9094
2.951	5.4292	0.96957	1.5017	4.4524
3.223	2.1678	4.1198	3.1795	2.7386
2.7846	2.6586	5.2953	4.4683	5.2
5.9322	2.9303	2.67	2.4306	3.6539
4.3405	3.555	4.77	4.698	2.59
0.4569	4.0828	2.9857	5.7852	4.669
3.7613	5.5942	1.7351	4.2128	3.7913

Annexe B

RÉSULTATS POUR LE GROUPE DES SEULS UTILISATEURS EN CONSIDÉRANT UNIQUEMENT LES ACCIDENTS AVEC BLESSÉS

Pour la répartition des appels, les résultats sont les mêmes que ceux obtenus pour le groupe des seuls utilisateurs avec tous les types d'accidents présentés au chapitre trois. Pour les accidents, nous fournissons le tableau permettant de statuer sur le modèle à conserver (voir tableau B.1). Nous avons aussi ajouté le tableau B.2 des estimations pour les paramètres impliqués dans le modèle conservé (trois modes) ainsi que la figure B.1 qui présente la courbe des données et des densités des modèles considérés pour l'ajustement.

TABLEAU B.1. Comparaison des différents modèles pour les données sur les accidents

Modèle	$-\widehat{H}_f(\underline{\theta})$	**AIC**	**BIC**
1 mode	-1,649	-4,649	-10,232
2 modes	-1,636	-7,636	-18,803
3 modes	**-1,623**	-10,623	-27,373
4 modes	-1,625	-13,625	-35,959
5 modes	-1,626	-16,626	-44,543

TABLEAU B.2. Estimation des paramètres d'un mélange de densités marginales à trois composantes pour les données sur les accidents

	Trois composantes				
Param.	**Val. initiale**	**Val. finale**	**Param.**	**Val. initiale**	**Val. finale**
p_1	0,270	0,220	ω_1	200,000	266,564
p_2	0,240	0,330	ω_2	150,000	265,781
p_3	0,490	0,450	ω_3	2,000	1,041
μ_1	2,000 (7h38)	2,142 (8h11)	μ_2	3,800 (14h31)	3,766 (14h23)
μ_3	4,900 (18h43)	4,675 (17h51)	κ	2,400	2,613

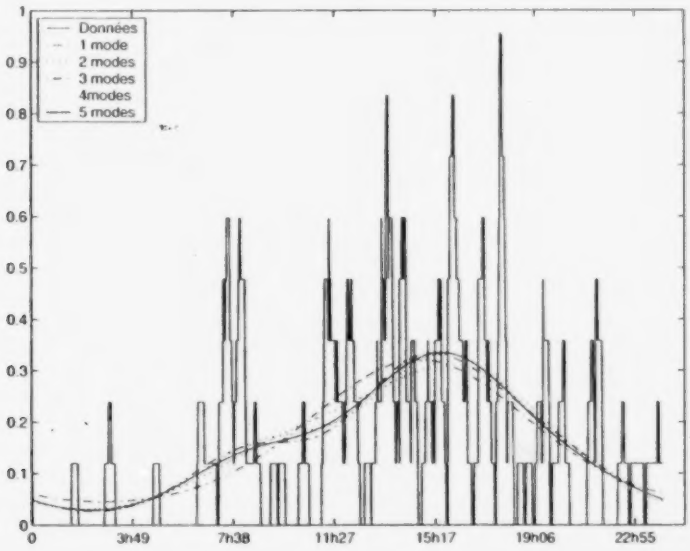

FIGURE B.1. Comparaison de la répartition des accidents selon l'heure de la journée et son approximation par un mélange à trois composantes.

Pour l'estimation des densités pour les probabilités marginales transformées, nous fournissons le tableau B.3 permettant de statuer sur le modèle à conserver pour les appels et le tableau B.5 pour les accidents. Nous avons aussi ajouté les

tableaux B.4 pour les appels et B.6 pour les accidents qui donnent les estimations pour les paramètres impliqués dans chacun des modèles conservés. Les figures B.2 et B.3 montrent l'histogramme des données transformées (voir équation (3.2.4)), pour les appels ($D_{(1)} = 0{,}010$ et $D_{(n)} = 0{,}011$) et pour les accidents ($D_{(1)} = 0{,}002$ et $D_{(n)} = 0{,}022$) respectivement, et les courbes estimées selon le nombre de composantes dans le mélange de lois.

TABLEAU B.3. Comparaison des différents modèles pour les probabilités transformées pour les appels

Modèle	$-\widehat{H}_f(\underset{\sim}{\theta})$	**AIC**	**BIC**
1 mode	1,254	-1,254	-1,76
2 modes	**4,177**	-1,177	-3,358

TABLEAU B.4. Estimation des paramètres d'un mélange de densités bêta à deux composantes pour les données sur les appels

Deux composantes					
Param	**Val. initiale**	**Val. finale**	**Param**	**Val. initiale**	**Val. finale**
p_1	0,419	0,429	α_2	9,730	9,709
p_2	0,581	0,571	β_1	3,226	2,256
α_1	2,178	1,695	β_2	1,064	1,419

TABLEAU B.5. Comparaison des différents modèles pour les probabilités transformées pour les accidents

Modèle	$-\widehat{H}_f(\underset{\sim}{\theta})$	**AIC**	**BIC**
1 mode	1,149	-1,149	-1,865
2 modes	3,263	-2,263	-9,273
3 modes	**12,053**	4,053	-0,004

Pour le groupe des seuls utilisateurs avec accidents avec blessés, 6360 individus participaient à l'étude. Nous avons comptabilisé un total de 128 accidents

TABLEAU B.6. Estimation des paramètres d'un mélange de densités bêta à deux composantes pour les données sur les appels

Deux composantes					
Param	**Val. initiale**	**Val. finale**	**Param**	**Val. initiale**	**Val. finale**
p_1	0,086	0,097	α_3	10,936	10,906
p_2	0,438	0,402	β_1	15,320	15,278
p_3	0,476	0,501	β_2	10,695	10,582
α_1	1,636	2,198	β_3	1,348	1,736
α_2	8,805	8,510			

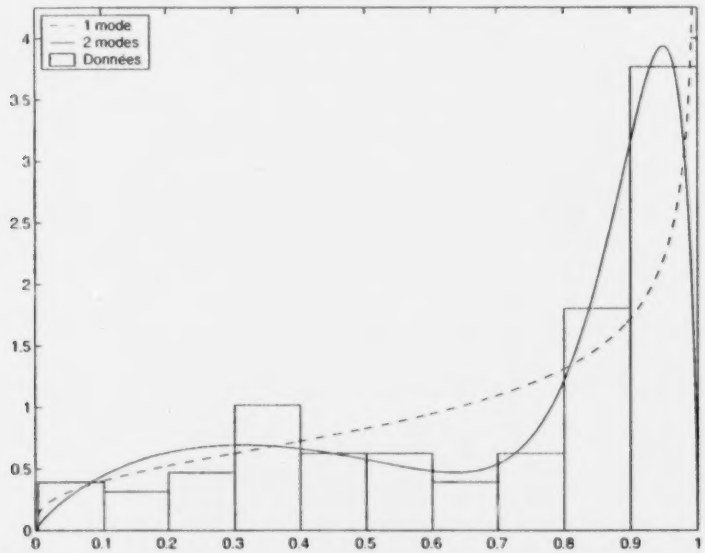

FIGURE B.2. Comparaison de la répartition des probabilités transformées d'appel et son approximation par un modèle à trois composantes dans un mélange de densités bêta

avec blessés faits par 124 personnes différentes.

Voici maintenant les différentes valeurs connues dans la reconstruction du tableau de contingence 2×2. Le nombre d'individus/minute dans ce cas est donné par $n = 95460$. Le nombre d'accidents/minute avec téléphone mobile est $n_{11} = 42$ si nous considérons le dernier appel fait dans l'intervalle de 15 minutes précédent le temps de l'accident, $n_{11}^{\min} = 26$ si nous considérons le temps d'appel minimum

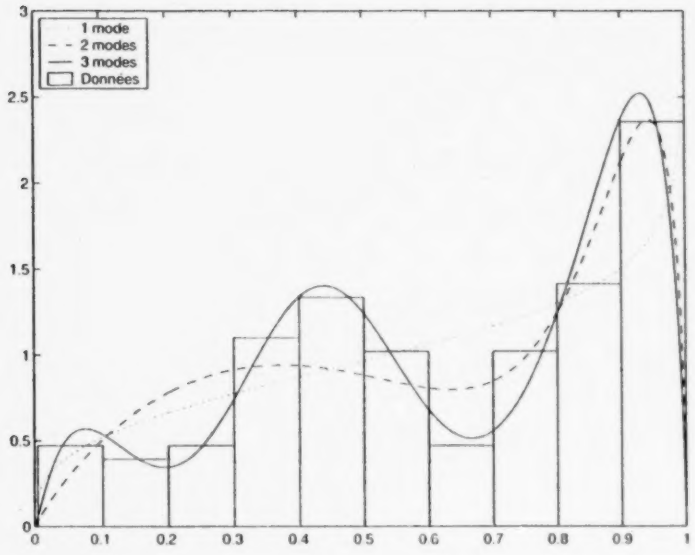

FIGURE B.3. Comparaison de la répartition des probabilités transformées d'accident et son approximation par un modèle à trois composantes dans un mélange de densités bêta

dans ce même intervalle et $n_{11}^{\max} = 48$ si nous considérons plutôt le temps d'appel maximum. Selon l'hypothèse considérée, nous obtenons $n_{21} = 1878$, $n_{21}^{\min} = 1894$ et $n_{21}^{\max} = 1872$. Finalement nous fournissons les résultats des estimations du risque relatif et du rapport de cotes ainsi que les régions α-crédibles correspondantes aux tableaux B.7 et B.8.

TABLEAU B.7. Estimation du rapport de cotes et du risque relatif par la méthode de Monte Carlo avec fonction d'importance pour les seuls utilisateurs pour les accidents avec blessés

Hypothèse	**Mesure**	**Estimateur**	**Précision**
n_{11}	Risque relatif	2,000	0,001
	Rapport de cotes	2,041	0,002
$n_{11}^{\min}$	Risque relatif	1,244	0,001
	Rapport de cotes	1,250	0,001
$n_{11}^{\max}$	Risque relatif	2,284	0,002
	Rapport de cotes	2,345	0,002

TABLEAU B.8. Comparaison des régions α-crédibles pour le risque relatif et le rapport de cotes selon l'hypothèse et de niveau $1 - \alpha = 0{,}95$.

Hypothèse	**Risque relatif**	**Rapport de cotes**
n_{11}	[1,884 ; 2,136]	[1,918 ; 2,041]
$n_{11}^{\min}$	[1,169 ; 1,323]	[1,173 ; 1,332]
$n_{11}^{\max}$	[2,148 ; 2,429]	[2,199 ; 2,502]

Annexe C

RÉSULTATS POUR LE GROUPE DES GRANDS UTILISATEURS EN CONSIDÉRANT TOUS LES TYPES D'ACCIDENTS

Pour la répartition des appels et des accidents, nous fournissons les tableaux permettant de statuer sur le modèle à conserver (voir tableaux C.1 et C.2). Nous avons aussi ajouté les tableaux C.3 et C.4 des estimations pour les paramètres impliqués dans chacun des modèles conservés ainsi que les figures C.1 et C.2 qui présentent la courbe des données et des densités pour les modèles considérés pour l'ajustement.

TABLEAU C.1. Comparaison des différents modèles pour les données sur les appels

Modèle	$-\widehat{H}_J(\underline{\theta})$	**AIC**	**BIC**
1 mode	-1,837	**-4,837**	**-28,515**
2 modes	-1,837	-7,837	-55,194
3 modes	-1,837	-10,837	-81,873

Pour l'estimation des densités pour les probabilités marginales transformées, nous fournissons le tableau C.5 permettant de statuer sur le modèle à conserver pour les appels et le tableau C.7 pour les accidents. Nous avons aussi ajouté les tableaux C.6 pour les appels et C.8 pour les accidents qui donnent les estimations pour les paramètres impliqués dans chacun des modèles conservés. Les

TABLEAU C.2. Comparaison des différents modèles pour les données sur les accidents

Modèle	$-\widehat{H}_f(\underset{\sim}{\theta})$	**AIC**	**BIC**
1 mode	-1,665	-4,665	-10,176
2 modes	-1,620	-7,620	-18,642
3 modes	**-1,617**	-10,617	-27,151
4 modes	-1,620	-13,620	-35,665
5 modes	-1,632	-16,632	-44,189

TABLEAU C.3. Estimation des paramètres de la densité unimodale pour les données sur les appels

Param.	**Val. initiale**	**Val. finale**
μ_1	3,738 (14h17)	3,740 (14h17)
ω_1	0,068	0,152
κ	0,990	1,000

TABLEAU C.4. Estimation des paramètres d'un mélange de densités marginales à trois composantes pour les données sur les accidents

	Trois composantes				
Param.	**Val. initiale**	**Val. finale**	**Param.**	**Val. initiale**	**Val. finale**
p_1	0,330	0,366	ω_1	25,000	93,062
p_2	0,330	0,303	ω_2	25,000	1,278
p_3	0,340	0,331	ω_3	1,000	1,186
μ_1	2,500 (9h33)	2,543 (4h43)	μ_2	4,000 (15h17)	3,949 (15h05)
μ_3	5,000 (19h06)	4,964 (18h58)	κ	2,400	3,083

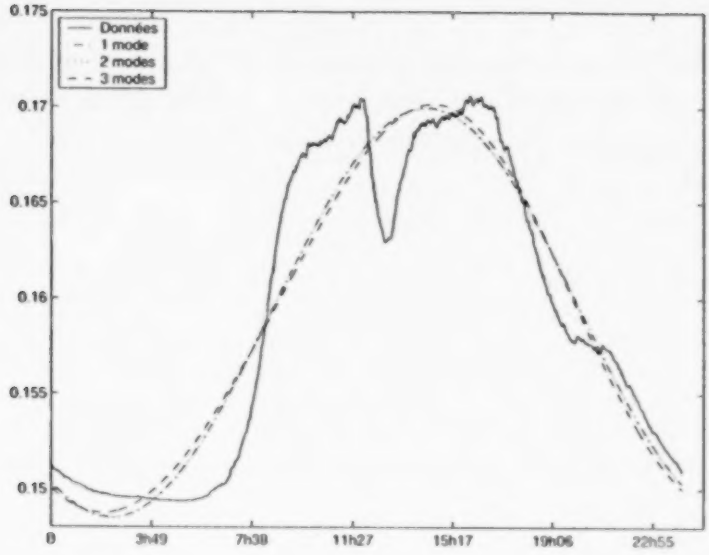

FIGURE C.1. Comparaison de la répartition des appels selon l'heure de la journée et son approximation par différents modèles.

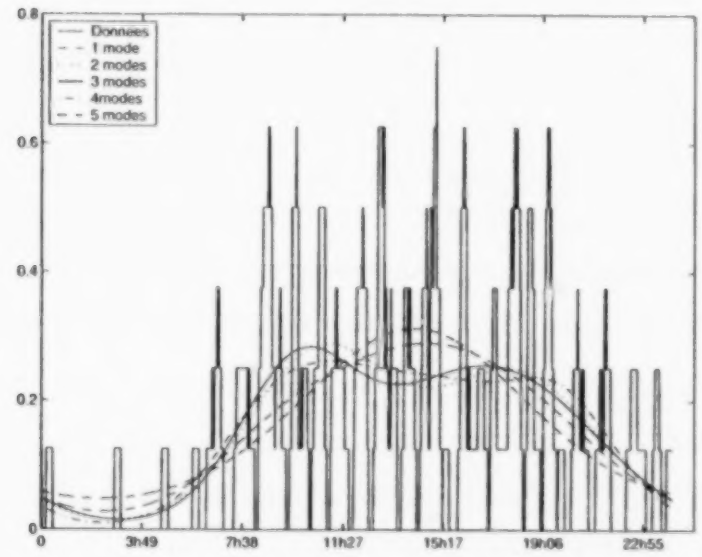

FIGURE C.2. Comparaison de la répartition des accidents selon l'heure de la journée et son approximation par différents modèles.

figures C.3 et C.4 montrent l'histogramme des données transformées (voir équation (3.2.4)), pour les appels ($D_{(1)} = 0{,}009$ et $D_{(n)} = 0{,}011$) et pour les accidents ($D_{(1)} = 9{,}981 \times 10^{-4}$ et $D_{(n)} = 0{,}019$) respectivement, et les courbes estimées

selon le nombre de composantes dans le mélange de lois.

TABLEAU C.5. Comparaison des différents modèles pour les probabilités transformées pour les appels

Modèle	$-\widehat{H}_f(\underline{\theta})$	AIC	BIC
1 mode	1,380	-1,380	-3,586
2 modes	7,234	2,234	-0,182
3 modes	**7,792**	-0,208	-4,072

TABLEAU C.6. Estimation des paramètres d'un mélange de densités bêta à trois composantes pour les données sur les appels

Trois composantes					
Param	**Val. initiale**	**Val. finale**	**Param**	**Val. initiale**	**Val. finale**
p_1	0,172	0,154	α_3	32,772	32,740
p_2	0,508	0,479	β_1	9,853	9,900
p_3	0,320	0,367	β_2	3,594	4,016
α_1	2,956	2,678	β_3	1,141	2,519
α_2	7,493	7,303			

TABLEAU C.7. Comparaison des différents modèles pour les probabilités transformées pour les accidents

Modèle	$-\widehat{H}_f(\underline{\theta})$	AIC	BIC
1 mode	1,386	-1,386	-3,581
2 modes	**5,037**	0,037	-2,379

Pour le groupe des grands utilisateurs avec tous les types d'accidents, 1597 individus participaient à l'étude. Nous avons comptabilisé un total de 122 accidents faits par 111 personnes différentes.

Voici maintenant les différentes valeurs connues dans la reconstruction du tableau de contingence 2×2. Le nombre d'individus/minute dans ce cas est donné

TABLEAU C.8. Estimation des paramètres d'un mélange de densités bêta à deux composantes pour les données sur les accidents

Deux composantes					
Param	**Val. initiale**	**Val. finale**	**Param**	**Val. initiale**	**Val. finale**
p_1	0,123	0,130	α_2	7,076	7,411
p_2	0,877	0,870	β_1	10,525	10,477
α_1	2,884	3,080	β_2	10,616	1,858

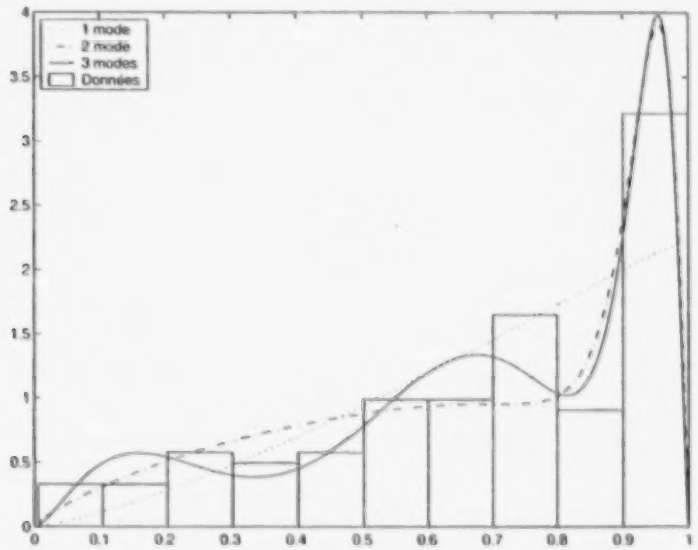

FIGURE C.3. Comparaison de la répartition des probabilités transformées d'appel et son approximation par différents modèles

par $n = 24105$. Le nombre d'accidents/minute avec téléphone mobile est $n_{11} = 48$ si nous considérons le dernier appel fait dans l'intervalle de 15 minutes précédent le temps de l'accident, $n_{11}^{\min} = 36$ si nous considérons le temps d'appel minimum dans ce même intervalle et $n_{11}^{\max} = 72$ si nous considérons plutôt le temps d'appel maximum. Selon l'hypothèse considérée, nous obtenons $n_{21} = 1782$, $n_{21}^{\min} = 1794$ et $n_{21}^{\max} = 1758$. Finalement nous fournissons les résultats des estimations du risque relatif et du rapport de cotes ainsi que les régions α-crédibles correspondantes aux tableaux C.9 et C.10.

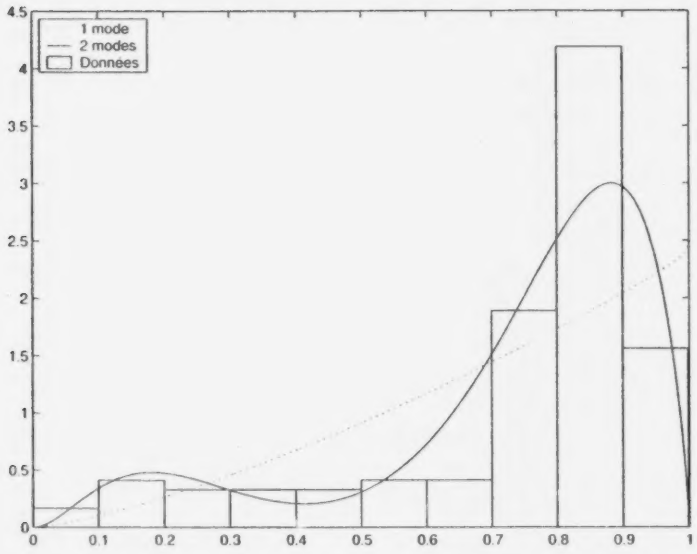

FIGURE C.4. Comparaison de la répartition des probabilités transformées d'accident et son approximation par différents modèles

TABLEAU C.9. Estimation du rapport de cotes et du risque relatif par la méthode de Monte Carlo avec fonction d'importance pour les grands utilisateurs pour tous les types d'accidents

Hypothèse	**Mesure**	**Estimateur**	**Précision**
n_{11}	Risque relatif	2,270	0,003
	Rapport de cotes	2,532	0,004
$n_{11}^{\min}$	Risque relatif	1,769	0,002
	Rapport de cotes	1,888	0,003
$n_{11}^{\max}$	Risque relatif	3,192	0,004
	Rapport de cotes	3,872	0,006

TABLEAU C.10. Comparaison des régions α-crédibles pour le risque relatif et le rapport de cotes selon l'hypothèse et de niveau $1 - \alpha = 0{,}95$.

Hypothèse	**Risque relatif**	**Rapport de cotes**
n_{11}	[2,034 ; 2,529]	[2,220 ; 2,886]
$n_{11}^{\min}$	[1,576 ; 1,995]	[1,654 ; 2,171]
$n_{11}^{\max}$	[3,872 ; 4,431]	[3,388 ; 4,431]

Annexe D

RÉSULTATS POUR LE GROUPE DES PETITS UTILISATEURS EN CONSIDÉRANT TOUS LES TYPES D'ACCIDENTS

Pour la répartition des appels et des accidents, nous fournissons les tableaux permettant de statuer sur le modèle à conserver (voir tableaux D.1 et D.2). Nous avons aussi ajouté les tableaux D.3 et D.4 des estimations pour les paramètres impliqués dans chacun des modèles conservés ainsi que les figures D.1 et D.2 qui présentent la courbe des données et des densités pour les modèles considérés pour l'ajustement.

TABLEAU D.1. Comparaison des différents modèles pour les données sur les appels

Modèle	$-\widehat{H}_f(\underline{\theta})$	AIC	BIC
1 mode	-1,832	**-4,832**	**-22,198**
2 modes	-1,832	-7,832	-42,564
3 modes	-1,832	-10,832	-62,930

Pour l'estimation des densités pour les probabilités marginales transformées, nous fournissons le tableau D.5 permettant de statuer sur le modèle à conserver pour les appels et le tableau D.7 pour les accidents. Nous avons aussi ajouté les tableaux D.6 pour les appels et D.8 pour les accidents qui donnent les estimations pour les paramètres impliqués dans chacun des modèles conservés. Les

TABLEAU D.2. Comparaison des différents modèles pour les données sur les accidents

Modèle	$-\widehat{H}_f(\underline{\theta})$	**AIC**	**BIC**
1 mode	-1,667	-4,667	-9,063
2 modes	-1,636	-7,636	-16,427
3 modes	-1,601	-10,601	-23,789
4 modes	**-1,576**	**-13,576**	**-31,160**
5 modes	-1,576	-16,576	-38,526

TABLEAU D.3. Estimation des paramètres de la densité unimodale pour les données sur les appels

Param.	**Val. initiale**	**Val. finale**
μ_1	4,002 (15h17)	3,999 (15h16)
ω_1	0,158	0,360
κ	0,990	1,000

figures D.3 et D.4 montrent l'histogramme des données transformées (voir équation (3.2.4)), pour les appels ($D_{(1)} = 0{,}009$ et $D_{(n)} = 0{,}012$) et pour les accidents ($D_{(1)} = 2{,}381 \times 10^{-5}$ et $D_{(n)} = 0{,}112$) respectivement, et les courbes estimées selon le nombre de composantes dans le mélange de lois.

Pour le groupe des petits utilisateurs avec tous les types d'accidents, 1892 individus participaient à l'étude. Nous avons comptabilisé un total de 58 accidents faits par 55 personnes différentes.

Voici maintenant les différentes valeurs connues dans la reconstruction du tableau de contingence 2×2. Le nombre d'individus/minute dans ce cas est donné par $n = 28425$. Le nombre d'accidents/minute avec téléphone mobile est $n_{11} = n_{11}^{\min} = n_{11}^{\max} = 8$. Peu importe l'hypothèse considérée, nous obtenons $n_{21} = 862$. Finalement nous fournissons les résultats des estimations du risque

TABLEAU D.4. Estimation des paramètres d'un mélange de densités marginales à quatre composantes pour les données sur les accidents

	Quatre composantes				
Param.	**Val. initiale**	**Val. finale**	**Param.**	**Val. initiale**	**Val. finale**
p_1	0,250	0,245	ω_1	25,000	33,502
p_2	0,250	0,235	ω_2	18,000	18,900
p_3	0,250	0,298	ω_3	20,000	26,802
p_4	0,250	0,222	ω_4	20,000	0,583
μ_1	2,000 (7h38)	2,362 (9h01)	μ_2	3,000 (11h27)	2,755 (10h31)
μ_3	4,000 (15h17)	4,188 (16h00)	μ_4	5,000 (19h06)	5,344 (20h25)
κ	3,500	4,305			

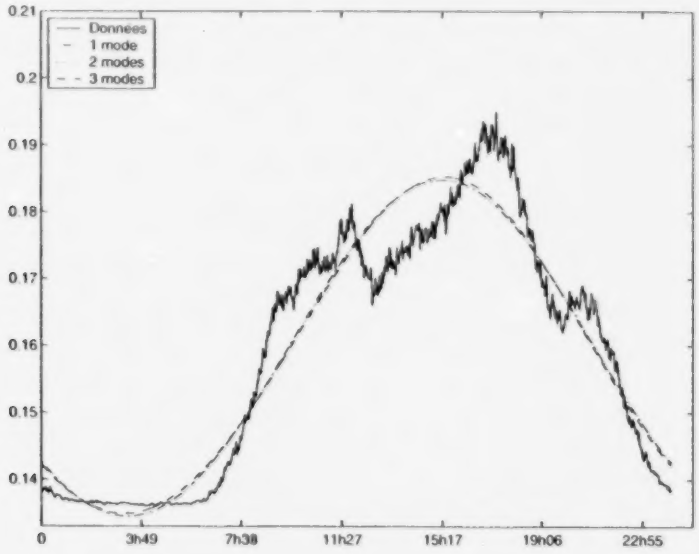

FIGURE D.1. Comparaison de la répartition des appels selon l'heure de la journée et son approximation par différents modèles.

relatif et du rapport de cotes ainsi que les régions α-crédibles correspondantes aux tableaux D.9 et D.10.

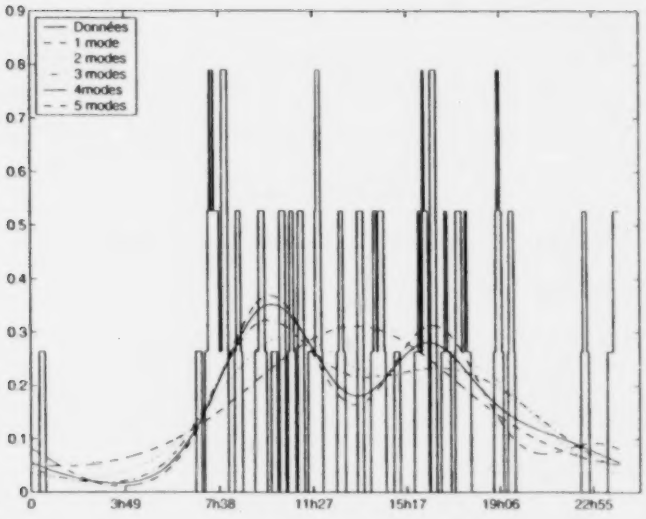

FIGURE D.2. Comparaison de la répartition des accidents selon l'heure de la journée et son approximation par différents modèles.

TABLEAU D.5. Comparaison des différents modèles pour les probabilités transformées pour les appels

Modèle	$-\widehat{H}_f(\underset{\sim}{\theta})$	**AIC**	**BIC**
1 mode	1,194	-1,194	-3,028
2 modes	9,581	4,581	2,025
3 modes	**13,984**	5,984	5,094

TABLEAU D.6. Estimation des paramètres d'un mélange de densités bêta à trois composantes pour les données sur les appels

Trois composantes					
Param	**Val. initiale**	**Val. finale**	**Param**	**Val. initiale**	**Val. finale**
p_1	0,310	0,308	α_3	45,037	45,006
p_2	0,379	0,349	β_1	18,953	18,900
p_3	0,310	0,343	β_2	3,925	4,279
α_1	6,926	7,111	β_3	1,527	2,966
α_2	9,863	9,746			

TABLEAU D.7. Comparaison des différents modèles pour les probabilités transformées pour les accidents

Modèle	$-\widehat{H}_f(\underset{\sim}{\theta})$	**AIC**	**BIC**
1 mode	3,000	1,000	0,778
2 modes	**3,981**	-1,019	-1,576

TABLEAU D.8. Estimation des paramètres d'un mélange de densités bêta à deux composantes pour les données sur les accidents

Deux composantes					
Param	**Val. initiale**	**Val. finale**	**Param**	**Val. initiale**	**Val. finale**
p_1	0,970	0,966	α_2	13,000	13,041
p_2	0,003	0,034	β_1	6,665	9,313
α_1	2,000	1,719	β_2	1,631	1,130

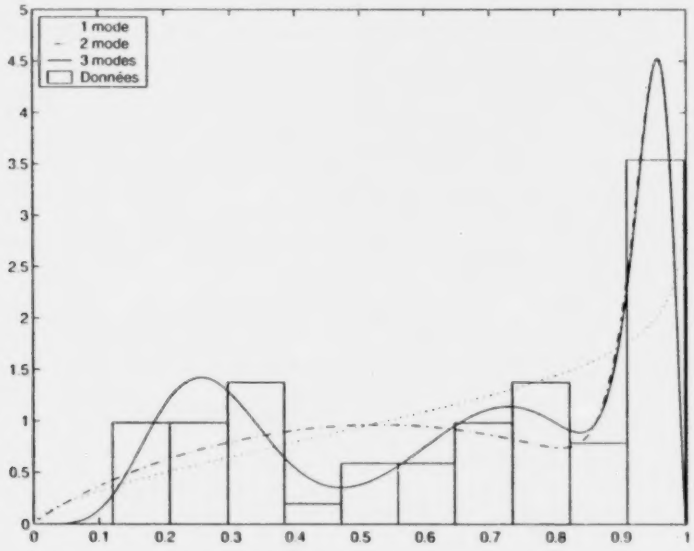

FIGURE D.3. Comparaison de la répartition des probabilités transformées d'appel et son approximation par différents modèles

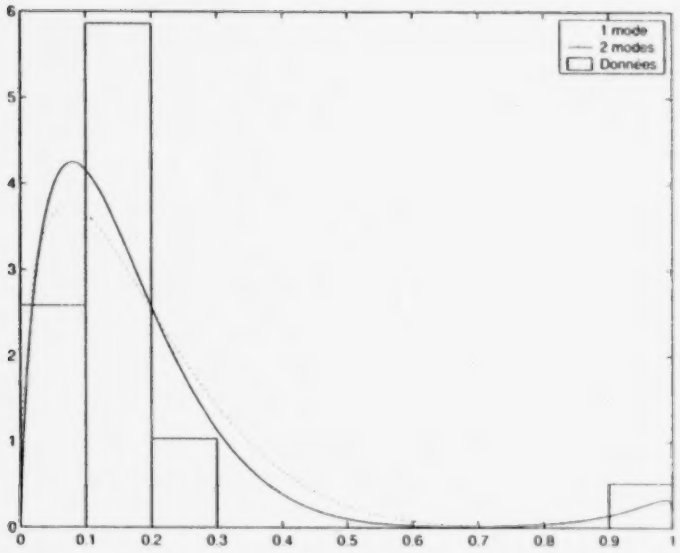

FIGURE D.4. Comparaison de la répartition des probabilités transformées d'accident et son approximation par différents modèles

TABLEAU D.9. Estimation du rapport de cotes et du risque relatif par la méthode de Monte Carlo avec fonction d'importance pour les petits utilisateurs pour tous les types d'accidents

	Mesure	**Estimateur**	**Précision**
n_{11}	Risque relatif	0,837	0,001
	Rapport de cotes	0,832	0,001

TABLEAU D.10. Région α-crédible pour le risque relatif et le rapport de cotes de niveau $1 - \alpha = 0{,}95$.

	Risque relatif	**Rapport de cotes**
	[0,736 ; 0,957]	[0,723 ; 0,956]

Annexe E

PROCÉDURE GÉNÉRALE

Pour chacun des jeux de données (appels et accidents), nous avons suivi la procédure qui suit. Les programmes informatiques qui ont été utilisés sont disponibles au bureau de Jean-François Angers.

1. Création du vecteur du nombre d'appels ou d'accidents pour chaque minute de la journée (appel.m ou accident.m).
2. Estimation des paramètres impliqués dans le modèle (loi marginale de VM) en fonction du nombre de composantes (Em4mod.m).
3. Choix du modèle à conserver selon les critères du maximum d'entropie, AIC et BIC (Em4mod.m).
4. Détermination des bornes d'intégration pour chacun des accidents (bornes.m).
5. Détermination de $D_{(1)}$ et $D_{(n)}$ (minmax.m).
6. Création de l'histogramme des probabilités transformées (histo.m).
7. Estimation des paramètres impliqués dans le modèle (densité bêta) en fonction du nombre de composantes (Emmod3.m).
8. Choix du modèle à conserver selon les critères du maximum d'entropie, AIC et BIC (Emmod3.m).
9. Application de la méthode de Monte Carlo avec fonction d'importance (esp.m, cst.m et mc.m).
10. Détermination des régions α-crédibles (RR.m et RC.m).

BIBLIOGRAPHIE

ABRAMOWITZ, M. ET STEGUN I.A.(1964), *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*, U.S. Government Printing Office, Washington.

AGRESTI, A.(1990), *Categorical Data Analysis*, Wiley, New York.

AIKAIKE, H.(1974), A new look at the statistical identification model, *IEEE Transactions on Automatic Control*, **19**, pp. 716-723.

ALM, H. ET NILSSON, L.(1995), The effects of a mobile telephone task on driver behaviour in a car following situation, *Accident Analysis and Prevention*, **27**, pp. 707-715.

BAKER, S.(1971), Digit preference in reported time of collision, *Accident Analysis and Prevention*, **3**, pp. 77-80.

BECKMAN, R.J. ET TIETJEN, G.L.(1978), Maximum likelihood estimation for the beta distribution, *Journal of Statistical Computation and Simulation*, **7**, pp. 253-258.

BERGER, J.O.(1985), *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, New York.

BURDEN, R.L. ET FAIRES, J.D.(1997), *Numerical Analysis*, Brooks/Cole, Pacific Grove.

CANADIAN WIRELESS TELECOMMUNICATIONS ASSOCIATION (CWTA)(1998), *Usage and Attitudes toward Wireless Communications in Canada*, OSI Technology Marketing Research Group.

DEMPSTER, A.P., LAIRD, N.M. ET RUBIN, D.B.(1977), Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society* (Série B), **39**, pp. 1-38.

FISHER, N.I.(1993), *Statistical Analysis of Circular Data*, Cambridge University Press, Cambridge.

FRASER, M.D., HSU Y. ET WALKER, J.J.(1981), Identifiability of finite mixtures of von Mises distributions, *The Annals of Statistics*, **9**, pp. 1130-1131.

FREDETTE, M.(1999), *Approche bayésienne pour estimer le rapport de cotes dans un tableau de contingence 2 × 2*, Mémoire de maîtrise, Département de mathématiques et de statistique, Université de Montréal.

GART, J.G. ET ZWEIFUL, R.Z.(1967), On the bias of various estimators of the logit and its variance with application to quantal bioassay, *Biometrika*, **54**, pp. 181-187.

GUMBEL, E.J.(1954), Applications of the circular normal distributions, *Journal of the American Statistical Association*, **49**, pp. 267-297.

HALDANE, J.B.S.(1955), The estimation and significance of the logarithm of a ratio of frequencies, *Annals of Human Genetics*, **20**, pp. 309-311.

JOHNSON, M.E.(1987), *Multivariate Statistical Simulation*, Wiley, New York.

JONES, T.A. ET JAMES, W.R.(1969), Analysis of bimodal orientation data, *Mathematical Geology*, **1**, pp. 129-135.

LABERGE-NADEAU, C., MAAG, U., BELLAVANCE, F., DESJARDINS, D., MESSIER, S., SAÏDI, A.(2001), *Les téléphones mobiles/cellulaires et le risque d'accidents*, Rapport de recherche CRT-2001-03, Centre de recherche sur les transports, Université de Montréal.

LEONARD, T. ET HSU, J.S.J.(1999), *Bayesian Methods : An Analysis for Statisticians and Interdisciplinary Researchers*, Cambridge University Press, Cambridge.

MARDIA, K.V. ET EL-ATOUM, A.A.M.(1976), Bayesian inference for the von Mises-Fisher Distribution, *Biometrika*, **63**, pp. 203-206.

MARDIA, K.V. ET SUTTON, T.W.(1975), On the modes of a mixture of two von Mises distributions, *Biometrika*, **62**, pp. 699-701.

MARITZ, J.(1970), *Empirical Bayes Methods*, Methuen, London.

MCKNIGHT, A.J. ET MCKNIGHT, A.S.(1993), The effects of cellular phone use upon driver attention, *Accident Analysis and Prevention*, **25**, pp. 259-265.

PEARSON, K.(1904), Mathematical contributions to the theory of evolution XIII : On the theory of contingency and its relation to association and normal correlation, *Biometric Series*, **1**.

RAIFFA, H. ET SCHLAIFER, R.(1961), *Applied Statistical Decision Theory*, MIT Press, Cambridge.

REDELMEIER, D.A. ET TIBSHIRANI, R.J.(1997), Cellular telephones and motor-vehicle collisions : some variations on matched-pairs analysis, *Canadian Journal of Statistics*, **25**, pp. 581-591.

ROBERT, C.P.(1994), *The Bayesian Choice : from Decision-Theoretic Foundations to Computational Implementation*, Springer, New York.

SCHLESSELMAN, J.J.(1974), Sample size requirements in cohort and case-control studies of disease, *Journal of Epidemiology*, **99**, pp. 381-384.

SCHWARZ, G.(1978), Estimating the dimension of a model, *The Annals of Statistics*, **6**, pp. 461-464.

STEIN, A.C., PARSEGHIAN, Z. ET ALLEN, R.W.(1987), A simulator study of the safety implications of cellular mobile phone use, *31st Annual Proceedings of the American Association for Automotive Medicine*, pp. 181-200.

TITTERINGTON, D., SMITH, A. ET MAKOV, U.(1985), *Statistical Analysis of Finite Mixture Distributions*, Wiley, New York.

YAKOWITZ, S.J. ET SPRAGINS, J.D.(1968), On the identifiability of finite mixtures, *The Annals of Mathematical Statistics*, **39**, pp. 209-214.